DOCUMENT RESUME

ED 303 163 IR 052 613

TITLE Annual Review of OCLC Research, June 1987-June 1988.
INSTITUTION OCLC Online Computer Library Center, Inc., Dublin, Ohio.
PUB DATE 88
NOTE 89p.; For the 1985-86 review, see ED 278 397.
PUB TYPE Reference Materials - Bibliographies (131)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Advisory Committees; Annual Reports; *Bibliographic Utilities; Database Management Systems; *Information Retrieval; Information Science; Library Statistics; *Management Information Systems; *Man Machine Systems; Program Descriptions; *Research and Development Centers; *Research Projects; Union Catalogs; User Needs (Information)
IDENTIFIERS Full Text Searching; *OCLC

ABSTRACT

The projects reviewed in this annual report of the Online Computer Library Center, Inc. (OCLC) focus on four strategic areas: (1) enhancing the use of the OCLC Online Union Catalog; (2) investigating full document storage, retrieval, and presentation; (3) strengthening the interface between patron and system; and (4) providing statistical information for improved database management. Introductory materials include brief statements by individual members of the research advisory committee and an overview of the OCLC research program. This overview briefly describes the activities in each of the four strategic areas and how they contribute to the OCLC mission. Individual reports are then presented for the following: 17 OCLC projects; 6 external and collaborative research projects; and 7 library and information science research grant program projects. Abstracts of seven presentations in the distinguished seminar series are also provided. The report concludes with descriptions of research programs sponsored by OCLC, i.e., the visiting scholar program, postdoctoral fellowships, and research assistants program; biographical sketches of OCLC staff; an annotated listing of some other OCLC publications; and a prepaid reader survey card. (CGD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED303163

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR052613

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Lois Yoakam

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The purpose or purposes for which this Corporation is formed are to establish, maintain and operate a computerized library network and to promote the evolution of library use, of libraries themselves, and of librarianship, and to provide processes and products for the benefit of library users and libraries, including such objectives as increasing availability of library resources to individual library patrons and reducing the rate-of-rise of library per-unit costs, all for the fundamental public purpose of furthering ease of access to and use of the ever-expanding body of world wide scientific, literary and educational knowledge and information.

# Annual Review of OCLC Research

July 1987–June 1988

1988 by OCLC Online Computer Library Center, Inc.

Printed in the United States of America

ISSN 0891-198X

OCLC is a registered trademark and MARC EASI Reference and Search Editor are trademarks of OCLC Online Computer Library Center, Inc. Apollo is a registered trademark of Apollo Computer, Inc. Apple is a registered trademark Apple Computer, Inc. Bliss is a registered trademark of Bliss Technologies, Inc. Crosstalk is a trademark of Great Wave Software. Crosstalk is a registered trademark of Digital Communications Associates, Inc. dBase III is a registered trademark of Ashton-Tate. EDSQuick is a trademark of Quantum Leap Office Systems. DIALOG is a registered trademark of DIALOG Information Services, Inc. EDX is a trademark of Emerging Technology Consultants, Inc. Hercules is a registered trademark of Hercules Computer Technology, Inc. Hypercard is a trademark of Apple Computer, Inc. IBM and IBM PC are registered trademarks of International Business Machines Corporation. Metafont is a trademark of Addison-Wesley Publishing Company. Postscript is a registered trademark of Adobe Systems, Inc. ProCite is a trademark of Personal Bibliographic Software, Inc. Pyramid Technology is a trademark of Pyramid Technology Corporation. Sun is a registered trademark of Sun Microsystems, Inc. Tandem is a trademark of Tandem Computers, Inc. TEX is a trademark of the American Mathematical Society. UNIX is a trademark of AT&T Bell Laboratories. VAX and MicroVAX are trademarks of Digital Equipment Corporation. WY-700 is a registered trademark of Wyse Technology. XEROX is a registered trademark of Xerox Corporation. X Window System is a trademark of the Massachusetts Institute of Technology.

# Contents

Letter of Introduction, v

Research Advisory Committee, vii

1. OCLC Project Reports, 1

**Advanced Interface Management,** Martin Dillon and Martha Gordon Lindeman, **3**
**Automated Document Structure Analysis,** Stuart L. Weibel, **5**
**Clustering Equivalent Bibliographic Records,** Elaine Svenonius, **6**
**Collection Analysis,** Edward T. O'Neill, **8**
**Collection Analysis System,** David J. Stephens, **10**
**Connectionist Models of Computation,** Stuart L. Weibel, **11**
**Database Quality,** Edward T. O'Neill, **13**
**Document Facsimile Transmission Studies,** John C. Handley, **14**
**Duplicate Detection and the "Species Problem,"** John A. Bunge, **15**
**Graph-Text,** Thomas B. Hickey, **17**
**Information Display,** Thomas B. Hickey, **18**
**Information Services Project,** Mark Crook, **18**
**Library of Congress Class M Subset:Music,** Jeanette M. Drone, **20**
**Managing Large Retrievals,** Chandra G. Prabha, **22**
**Nonfiction Book Use,** Chandra G. Prabha, **24**
**Scholars Cross-Reference System,** Paul B. Kantor, **26**
**Selecting Bibliographic Records,** John A. Bunge, **27**

2. External and Collaborative Research, 29

**Automatic Phrase Construction for the Representation of Text Content,** Gerard Salton, **30**
**Extensions to the Advanced Interface Management Project,** Niall Teskey, **31**
**Increasing the Accessibility of Library of Congress Subject Headings in Online Bibliographic Systems,** Karen Markey and Diane Vizine-Goetz, **32**
**Latvian Bibliographic Records in the Online Union Catalog,** Inese Auzina Smith, **34**
**Mercury: An Electronic Library,** William Y. Arms et al., **36**
**Term Weighting in Document Retrieval,** Clement T. Yu and Hirotaka Mizuno, **37**

3. Library and Information Science Research Grant Program, 41

**Analytical Study of Bibliographic Data on a Title Page,** Billie Grace Herring, **42**
**Applying the Revised MARC Format for Three-Dimensional Objects in Museums,** Esther G. Bierbaum, **43**
**Converting English Subject Headings into Spanish,** Emilia Bernal-Rosa, **45**
**Evaluating Subject Collections,** Howard D. White, **46**
**Measurement of Subject Scatter in the Superintendent of Documents Classification,** Lee Shiflett, **49**
**Subject Representations in Monographic Records,** Pamela Reekes McKirdy, **51**
**Variations in Personal and Corporate Names in the OCLC Online Union Catalog,** Arlene G. Taylor, **53**

4. Distinguished Seminar Series, 55

**Sandra A. Mamrak:** Software Architecture to Support Data Translation, **56**
**Toni Carbo Bearman:** Changes in the Life Cycle of Information, **57**
**W. David Penniman:** Tomorrow's Library Today, **59**
**Nicholas J. Belkin:** On the Nature and Function of Explanation in Intelligent Information Retrieval, **61**
**Michael Lesk:** Automatic Sense Disambiguation, **62**
**William S. Cooper:** Probability Theory in Information Retrieval, **64**
**W. Bruce Croft:** Information Retrieval by Plausible Inference, **65**

Programs, 67

OCLC Staff, 70

Other OCLC Publications, 76

The *Annual Review of OCLC Research* is our report to OCLC members and the library and information science community of OCLC's research activities. We are eager to introduce you to the multifaceted OCLC research environment as well as to chronicle our research activities and summarize our important findings. We hope that this review will help to stimulate further creative research by you or others and that some of this will be cooperative research with OCLC.

The OCLC Office of Research is one of the larger centers of research devoted solely to solving the problems related to library and information science. OCLC's research activities focus on four major areas: enhancing the use of the OCLC Online Union Catalog; investigating full document storage, retrieval, and presentation; strengthening the interface between patron and information system; and providing statistical information for improved database management.

The problems tackled by OCLC research scientists are aimed at both immediate and long-term solutions for libraries today and for the libraries of the future. Whether providing a method to extract Online Union Catalog records for use in CAT CD450 (OCLC's compact disc cataloging system), exploring a statistically based collection analysis system, or working on user interfaces for future systems, OCLC research focuses on the application of experimental and theoretical results.

OCLC maintains commitment to continuing fruitful research. The natural outcome of this research is products and services for libraries, information professionals, and information users. The Offices of Research and Technology Assessment and the Development Division are administratively grouped to ensure coordination among these important OCLC capabilities.

OCLC research extends beyond the Center in Dublin, Ohio, by supporting research projects outside of OCLC and collaborating on joint projects. Thus, the impact of OCLC research is potentially important throughout the research community. This also helps to enhance the total impact of OCLC's research activities.

The future of library and information services continues to be exciting, challenging, fraught with significant research problems, and rich in potential for broad ranging societal impact. We hope this *Annual Review* will introduce you to the challenges, prompt you to consider the solutions, and encourage you to share the optimism of OCLC research.

Rowland C. W. Brown
President and Chief Executive Officer



Michael J. McGill
Vice President
Research, Technology Assessment, and Development

# The Research Advisory Committee

The Research Advisory Committee guides and evaluates the activities of the OCLC Office of Research. By providing valuable input derived from practical experience and research into the problems of library and information science, the committee helps set the course of OCLC research activities. From its inception the committee has encouraged freedom of intellectual pursuit without straying from the goals of OCLC: to provide computerized processes, products, and services for libraries and their users.

Committee members are appointed to a three-year term; members who served during 1987–88 and incoming members for 1988–89 appear below.

Robert M. Hayes

Activities of the OCLC Office of Research cover the full spectrum, from high-powered workstations to user studies and issues of library management, from high tech to the library user. This is as it should be; the OCLC research agenda seems balanced. I think this reflects the role of the Research Advisory Committee as a sounding board and reactive panel providing input to both the Office of Research and OCLC management."

"The discovery and application of new technology by OCLC have their place and are properly focused on advancing library and information services and products directed at the needs of library users and managers."

Robert M. Hayes is Dean, Graduate School of Library and Information Studies, University of California, Los Angeles (retiring member).

Brian Perry

This year marks the end of my three-year appointment as a member of the Research Advisory Committee, and I feel that there will be a great gap in my life! It has been both an honor and a pleasure to be associated with the work of the Office of Research, which I honestly consider to be one of the foremost research teams in the library and information science world."

"OCLC can justly be proud of the excellent researchers whom it has attracted, and it should be congratulated on the facilities it has made available to the Office and on the fact that, even in these difficult times, it has had the vision to support at least some research the results of which are not capable of immediate commercial exploitation, but which expand knowledge and understanding of this vital area of information activity."

Brian Perry is Director, Research and Development Department, The British Library (retiring member).

Brian L. Hawkins

The Research Advisory Committee facilitates dialog that is beneficial to OCLC and the broader higher education community. There is an increasing intertwining of technological innovation and traditional library services. The two worlds—computers and libraries—are becoming one."

"Full-text databases and compact disc storage will provide alternative access methods for libraries and their users. But technology should not be the only driving force in future library development. Libraries must express the pros and cons of technology and make known their needs and the needs of their users."

Brian L. Hawkins is Vice President for Computing and Information Services, Brown University (continuing member).

Mervin E. Muller

"Communication on the Research Advisory Committee is really a two-way street, a give and take that adds dimension to OCLC and the research community at large."

"My interests as a professional statistician focus on examining data to discover a profile of users and their needs. Statistical analyses help us understand how data and information systems are used, how technology affects users, and how users' needs should direct technological development.

"Research of this sort generates products that boost the productivity of users, whether libraries or patrons. Future research questions at OCLC should include what information users need and in what form, how they will use it, and what is the best interface to the information."

Mervin E. Muller is Chairman and Robert M. Critchfield Professor of Engineering, Computer and Information Science Department, and Professor, Department of Statistics, The Ohio State University (continuing member).

Joseph H. Howard

"I am excited about new and emerging technologies for libraries that will change both the ways they operate and the ways they are used. But libraries should also be proactive to tell database producers what kinds of information systems and products they need. Technological solutions must be affordable for libraries in light of constrained and sometimes diminishing financial resources."

"From the viewpoint of a national library, the possibility of establishing national repositories of machine-readable files is exciting. I am assuming that standards such as SGML will be adopted and used by publishers and database providers, eliminating the need for multiple translations. OCLC is a leader in this area."

Joseph H. Howard is Director, National Agricultural Library (incoming member).

Karen Sparck Jones

"OCLC's enterprises have influenced the library and information community in many different direct and indirect ways. OCLC's history and status give it a splendid opportunity for productive research."

"Its current research activities are usefully and interestingly varied. But I think there is a lot to be done in at least three relevant areas: making full-text browsing and extended reading at terminals convenient and agreeable, exploiting machine-readable dictionaries and similar reference tools to meet a range of needs, and developing and implementing sensible ideas on intelligent interfaces. All of these lines are relevant to the more comprehensive information systems of the future.

"I look forward to interacting with OCLC's research through the advisory committee."

Karen Sparck Jones is Assistant Director of Research, Computer Laboratory, University of Cambridge (incoming member).

# 1

# OCLC Project Reports

The Office of Research explores new services and products for OCLC and undertakes to solve problems encountered by OCLC and its membership in meeting the needs of library patrons. A secondary purpose is to provide research leadership to the library and information science community. The project mix at any one time, and the flow of projects over time, are planned to act together to advance the strategic aims set forth in the OCLC Charter.

At present, the Office is focused on four strategic areas: (1) enhancing the use of the OCLC® Online Union Catalog; (2) investigating full document storage, retrieval and presentation; (3) strengthening the interface between patron and system; and (4) providing statistical information for improved database management. A brief description of these activities and how they contribute to the OCLC mission follows.

## Enhanced Use of the OCLC Online Union Catalog

One of OCLC's major assets is the Online Union Catalog, and its usefulness depends on its size, the type of records it contains, their quality, and the tools provided to users. Many of the projects carried out in the Office of Research have as their objective enhancing the use of the OCLC database. The Database Quality and Duplicate Detection projects join many earlier projects in their focus on database quality. These projects have a threefold aim: to accurately depict the state of affairs of the database with respect to errors; to develop and apply techniques to reduce them; and to investigate long-term techniques for online database quality services. Additional projects also contribute to these goals. ▶

The Library of Congress Class M Subset project combines research into database quality in a specialized area with the goal of designing a database especially for reference services. The LC Subject Heading Analysis project similarly combines the goals of reference and improving the quality of the database, but centers on subject headings.

### Full Document Storage, Retrieval and Presentation

A number of projects in the Office of Research are centered on problems in the creation of document databases for use in full-text retrieval systems. Much of this activity is carried out in collaboration with Carnegie Mellon University (CMU) in our jointly sponsored Mercury project to establish an electronic library for CMU students and faculty. The library will serve as an experimental laboratory and test bed for research ideas and allow OCLC to explore issues that arise in the electronic delivery of full documents.

The core technology used by the Mercury project for the display of text and graphics was developed in the Graph-Text project. The Information Display project continues to investigate problems in the display of technical information, refining what we learned in the Graph-Text project and establishing the basis for the next generation of display devices. A second project, Automated Document Structure Analysis, focuses on the digitization of documents and the use of typographical clues to infer document structure. This research helps bridge the gap between paper information and structured electronic information by developing systems that transform paper images to a full-fledged Graph-Text database.

These two major projects join with teams from Carnegie Mellon University in the first collaborative effort between a major information provider and a university to create an electronic library (see section 2, page 36).

### Strengthening the Interface

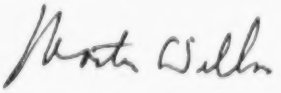
As OCLC extends its product lines and includes more services for the library patron, problems of interface design, functionality, and consistency assume increasing significance. The Advanced Interface Management project, with its particular emphasis on interfaces to full-text systems, deals with questions in this area. Our approach uses rapid prototyping of interface ideas within a software environment developed at OCLC for this purpose and probing analyses of what users really want. A second project, Managing Large Retrievals, focuses on patron use of the Online Union Catalog for general reference. Its objective is to develop tools and techniques that allow library patrons to carry out efficient subject searches using the OCLC database as a resource.

### Statistical Information Services

Increasingly, the Office of Research is called upon to support other units of OCLC when problems arise that require research for their solution, either in day-to-day operations or in support of development efforts. Recent examples include defining techniques for selecting an optimum subset from the Online Union Catalog for the CAT CD450 system (a compact disc cataloging system); providing statistical data for the Online Union Catalog, especially for cases where standard techniques are inadequate (estimating the number of titles that are duplicates, for example); database quality studies, especially gaining an accurate characterization of current database quality with respect to errors and duplicates; and developing cost models for products and services. These efforts are all handled under one project, Information Services.

These areas do not exhaust Office of Research projects. They do, however, identify the current priorities of the Office of Research and its staff.

One last activity of the Office of Research should be mentioned here, the Library and Information Science Research Grant program (LISRG). This program was established to exert an influence on the direction and quality of research in schools of library and information science by encouraging researchers to solve the kinds of practical cataloging and database problems that OCLC and its members face. Preliminary reports are presented in section 3, page 41.



Martin Dillon, Director
Office of Research

# Advanced Interface Management

**Project Managers:** Martin Dillon, Director, Office of Research; Martha Gordon Lindeman, Research Scientist

**Abstract**
This project is developing a standardized, highly portable environment for rapidly prototyping interfaces between users and online retrieval systems and for exploring issues in the design of interfaces to databases containing the full text of documents.

The Advanced Interface Management (AIM) project originally grew out of the need to prototype interfaces for retrieval systems. Due to an increasing emphasis on the retrieval of the full text of documents, the project has concentrated on defining the issues involved in full-text retrieval. The primary objective is to continue this generalization of the project, perhaps even beyond information retrieval, by developing a highly portable environment for rapid prototyping that exploits standardized tools.

We are using two approaches to achieve this goal: (1) testing and incorporating tools such as the X Window System™ into an environment that allows designers to prototype interfaces that may have several concurrent independent processes in multiple windows, and (2) conducting research that provides information relevant to full-text interface design. These two approaches are being integrated into the development of a research tool that allows us to design and quickly prototype the interfaces necessary for system-based experimentation on full-text retrieval in particular and interface design in general.

## Prototyping Environment

The prototyping environment uses a specially designed programming language called "D." The D language provides interface designers with development capabilities such as windowing (the ability to simultaneously display multiple processes and/or files on a single screen), processing user requests from several input devices, and displaying graphics, icons, and text. Interfaces designed with D can range from simple command-driven interfaces that have only one window to complex multitasking interfaces that have many windows with independent processes and numerous input devices.

A sample program segment from the current D language is shown in figure 1.1. The segment could occur at the beginning of a D program; it defines a display window (window 1), an associated icon that can be selected by the user to open the window, and an error window. By allowing the interface designer to define and manipulate objects at the window level, the designer can focus on interface design rather than low-level programming details.

We intend to generalize the conceptual base of the D language to support new ideas and research results in interface design. In addition, standardization and consistency have become major issues in interface design, with numerous vendors and organizations defining guidelines and standards. Our prototyping tool will make it easier for designers to conform to these guidelines and standards, as well as to explore ways of using them appropriately.

```
GLOBAL
        display_cursor = on
        special_char = (char = label = dumpscreen)
        keyboard_echo = on
end
STATE_INPUT STANDARD
        end_of_accept_character = "
        end_of_region_character = functionkey_right_14
        special_char = (char = functionkey_right_1,label = exit_conf_
        accept)
        mouse_left_down = (function = select,format = "[] ")
        mouse_right_down = (function = pulldown,format = "CHANGE")
        pulldown_menu_item = (format = "EXIT",label = exit_conf_accept)
Keyboard:
        window_var = window_1
        border_width = 3
        background = shaded
        width = 1
        height = 1
end
Icon:
        window_var = window_1
        border_width = 3
        background = shaded
        width = 35
        height = 74
        prompt = window_1_icon
end
end
WINDOW window_error:
        x_coordinate_start_value = 30
        y_coordinate_start_value = 20
        window_height = 5
        window_width = 25
        border = 3
        background = shaded
end
WINDOW window_1:
        x_coordinate_start_value = 0
        y_coordinate_start_value = 6
        window_height = 35
        window_width = 74
end
```

**Figure 1.1** D Language Program Segment

## Full-Text Interfaces

OCLC has begun to explore the potentialities of full-text databases. The provision of full-text databases is a new field; little is known about how people use printed documents and what they will want to do with full-text systems. We have collected interview data from university scholars to determine how they currently use printed documents, specifically reference books, monographs, journals, and journal articles. These data provide information about how scholars would use full-text retrieval systems to meet their professional responsibilities.

This research is now being extended to include other methodologies and members of other populations. Methodologies will include questionnaires, verbal protocols from

people "thinking aloud" while solving assigned problems using printed documents, and observation of problem solving. Populations will include users of some electronic mail networks, university students, and perhaps users of libraries. The results of this research will allow us to rank order the functionalities being considered for inclusion in a full-text retrieval system.

Figure 1.2 shows a sample screen from the first full-text prototype. Novices and experts can easily learn and use the system. Users may navigate through a book using mouse keys to display sequential pages. Selecting various icons or page numbers displays the desired page. Users may also "bookmark" pages and return to any bookmarked page by selecting an icon with a single press of a mouse key.

A more recent prototype that simulates a small personal computer (80-character by 24-line screen with mouse input) was developed to see if similar functionality could be developed within that set of constraints. We also added other functionality such as user-defined hypertext links. Users can link one section of the text to another and move between sections by selecting the link.

Because full-text access is such a new field, numerous issues require rethinking decisions made for earlier bibliographic systems. For example, one issue is how to organize documents within the database. This issue raises questions such as should an entire book or each of its paragraphs be treated as individual objects? We have analyzed the components of books, serials, and journal articles based on the Standard Generalized Markup Language (SGML) used by the Association of American Publishers. From this analysis, we developed a hierarchical structure that allows us to fit any book or article within that framework.

We intend to continue exploring interface issues in full-text retrieval, particularly within the context of the D-language prototyping environment. This includes issues of database representation, search and retrieval capabilities, and screen display.

Project staff: John R. Bonneau, Programmer/Analyst; J. Charles Crabb, Research Assistant; Kym Pocius, Research Assistant; Mahendra A. Ramachandran, Research Assistant; Anthony M. Repka, Senior Programmer/Analyst

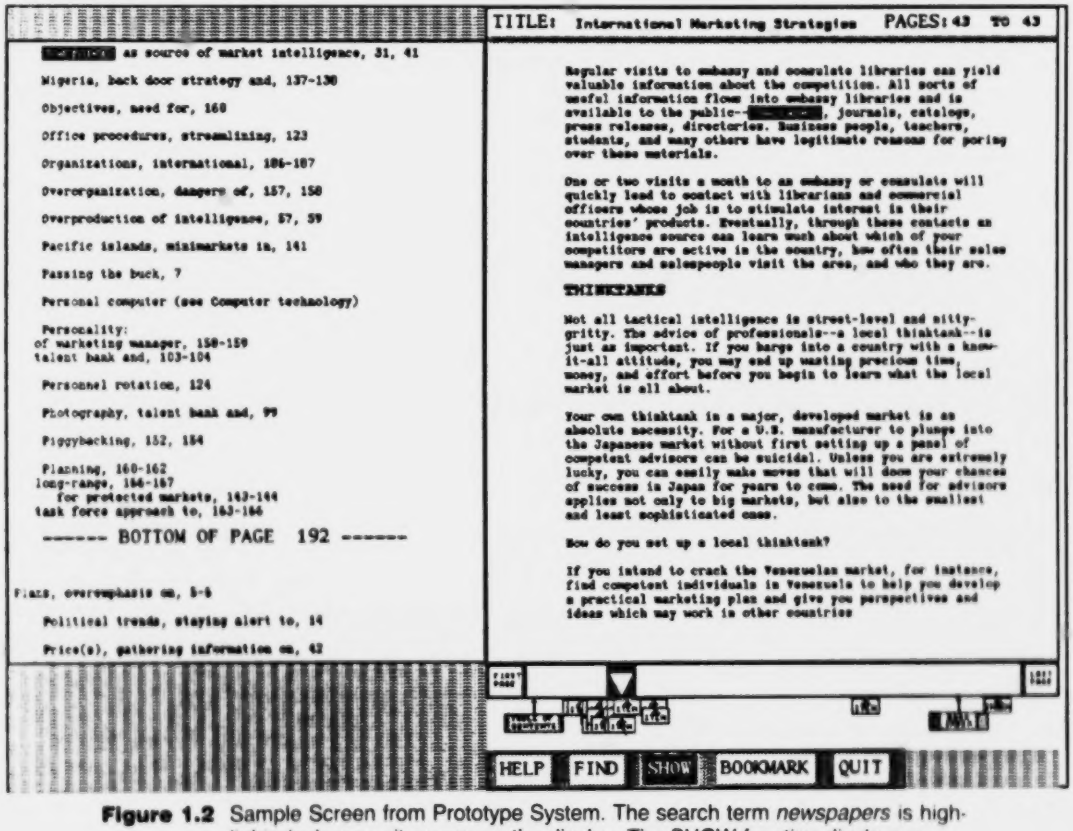

TITLE: International Marketing Strategies PAGES: 43 TO 43
as source of market intelligence, 31, 41
Nigeria, back door strategy and, 137-138
Objectives, need for, 160
Office procedures, streamlining, 123
Organizations, international, 186-187
Overorganization, dangers of, 157, 158
Overproduction of intelligence, 57, 59
Pacific islands, minimarkets in, 141
Passing the buck, 7
Personal computer (see Computer technology)
Personality:
of marketing manager, 158-159
talent bank and, 103-104
Personnel rotation, 124
Photography, talent bank and, 99
Piggybacking, 152, 154
Planning, 160-162
long-range, 166-167
for protected markets, 143-144
task force approach to, 163-166
------- BOTTOM OF PAGE 192 -------
Plans, overemphasis on, 5-6
Political trends, staying alert to, 14
Price(s), gathering information on, 42
Regular visits to embassy and consulate libraries can yield valuable information about the competition. All sorts of useful information flows into embassy libraries and is available to the public--[redacted], journals, catalogs, press releases, directories. Business people, teachers, students, and many others have legitimate reasons for poring over these materials.
One or two visits a month to an embassy or consulate will quickly lead to contact with librarians and commercial officers whose job is to stimulate interest in their countries' products. Eventually, through these contacts an intelligence source can learn much about which of your competitors are active in the country, how often their sales managers and salespeople visit the area, and who they are.
THINKTANKS
Not all tactical intelligence is street-level and nitty-gritty. The advice of professionals--a local thinktank--is just as important. If you barge into a country with a know-it-all attitude, you may end up wasting precious time, money, and effort before you begin to learn what the local market is all about.
Your own thinktank in a major, developed market is an absolute necessity. For a U.S. manufacturer to plunge into the Japanese market without first setting up a panel of competent advisors can be suicidal. Unless you are extremely lucky, you can easily make moves that will doom your chances of success in Japan for years to come. The need for advisors applies not only to big markets, but also to the smallest and least sophisticated ones.
How do you set up a local thinktank?
If you intend to crack the Venezuelan market, for instance, find competent individuals in Venezuela to help you develop a practical marketing plan and give you perspectives and ideas which may work in other countries
HELP
FIND
SHOW
BOOKMARK
QUIT

**Figure 1.2** Sample Screen from Prototype System. The search term *newspapers* is highlighted wherever it occurs on the display. The SHOW function displays a page from the index appearing in the left window. The mouse cursor (arrow) points to the icon used to display the text appearing in the right window.

# Automated Document Structure Analysis

**Project Manager:** Stuart L. Weibel, Research Scientist

**Abstract**
This project studies integration of state-of-the-art optical character recognition (OCR) capabilities into an automated document structure analysis system. OCR technology facilitates incorporation of documents into well-structured retrieval systems with minimum human intervention.

Identification and markup of bibliographic elements (e.g., title, author, publishing agency) and document structures (e.g., abstracts, text, graphs) are necessary to facilitate access and display of these elements in electronic information systems. Such systems have revolutionized how we gain access to information. Online access to bibliographic databases facilitates the use of library collections and promotes development of a world library—access to collections of networks of libraries. Demand for access to full-text databases is expected to grow in coming years, bringing the resources of libraries to the desktop workstations of scholars and researchers.

Incorporating full-text databases into effective retrieval systems requires tagging the text with a markup language to provide handles used to retrieve and display document components. The Standard Generalized Markup Language (SGML) predominates as a standard in this field.

Newly published materials are often available in machine-readable form, but publishers' typographic files use different tagging schemes, each of which requires unique decoding software. In addition, tapes often do not reflect last-pass editing in the publishing process. Therefore it is often better to obtain machine-readable files using optical character recognition (OCR) systems.

OCR represents a promising way to capture printed documents, but there are presently a number of uncertainties concerning the degree of editing and other processing necessary to economically create high-quality databases. Image processing techniques can "clean-up" a document prior to OCR processing and reduce the need for extensive postprocessing correction.

## Development Strategy

The system in development is a "blackboard" architecture expert system, a collection of expert subsystems applied to complex, multilevel problems such as document processing. The subsystems are discrete rule systems that can assert observations in a global data area known as the blackboard. For example, our system will have subsystems that "know" about document elements such as title pages, personal name formats, paragraph formatting, and bibliographic fields.

A scheduler module controls application of these subsystems. At any given time the state of the system (hypotheses about the structure of a document) is reflected in a global data structure (the blackboard) that all subsystems and the scheduler can read. The scheduler is essentially a higher-level rule system that tracks the hypotheses on the blackboard and schedules the activity of the subsystems appropriately.

The choice of a blackboard system is based upon the nature of document structure processing. Many disparate kinds of knowledge at different conceptual levels must be brought into play, from bit-level analysis to natural language processing. The modularity and extensibility of this architecture allow the system to grow gracefully as the need for additional subsystems develops.

Previous research confirmed that layout and keyword features can be useful for automatic structural analysis. Identifying these features can be done by prototyping a system to capture the underlying document architecture. Rules relate this architecture to relevant retrieval fields (e.g., authors, titles, publishing agency) and intrinsic document components such as text, abstracts, tables, indexes, tables of contents, graphs, and figures.

## Applications

We believe that our system will be useful in the following document processing and database building problems:

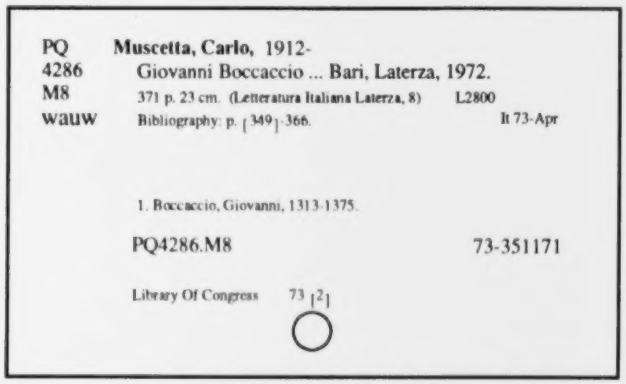
1. **Automated cataloging.** Our previous studies and those of others suggest significant opportunities to automate some measure of descriptive cataloging tasks.
2. **Retrospective conversion of catalog cards.** We expect that the process of converting catalog cards to MARC records can be facilitated substantially by application of OCR technology and expert systems.
3. **Scholar's personal databases.** The ability to scan documents and have them cataloged and indexed automatically in a personal database would be a boon to scholars.
4. **Archiving and preservation.** A system such as we envision would be a powerful tool to support institutional document management and the preservation of content where the "publisher's package" is of little importance.

## Preliminary Results

Among the problems encountered with document processing is automatically distinguishing among blocks of text, graphics (e.g., line drawings, halftones, and graphics of various types), and noise. Typically OCR devices attempt to interpret everything in a document image as text. We circumvent this problem by preprocessing images to segregate identifiable text blocks from nontext components in the image.

Preliminary experiments based on known image processing algorithms (Wong, Casey, and Wahl 1982) segment fields of text on catalog cards and distinguish them from noise such as smudges, stray marks or the rod hole.

Ninety-six catalog cards were optically scanned and the segmentation algorithm was applied to identify the bounding boxes of objects on the cards. One-third of the cards was used to develop a training set; the remaining cards were used to test the efficacy of the discriminant procedure. Each text object was identified manually; remaining objects were classified as noise. A typical card had about 15 text objects and 500 noise objects (figure 1.3). Statistical attributes were derived for each object used as input for a statistical classification procedure known as Classification and Regression Trees (CART) (Breiman, et al. 1984).



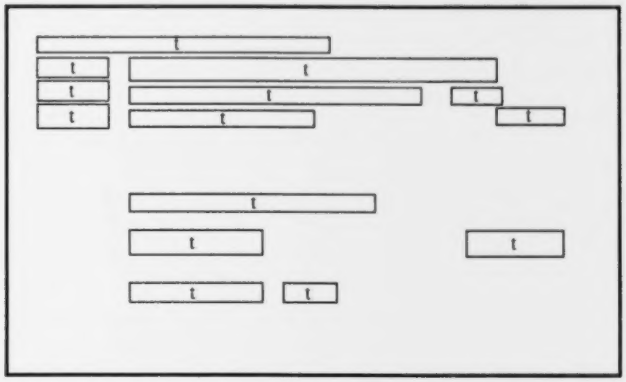



**Figure 1.3** Catalog Card and Text Objects. A typical catalog card image and the corresponding text bounding boxes identified in the card segmentation experiment.

The CART procedure determines the statistical profiles of the objects for the training set; these profiles are applied to the test set. A confusion matrix illustrates the high degree of accuracy achieved with this small sample (table 1.1).

By preprocessing a card image in this fashion, it should be possible to improve OCR performance significantly (the OCR device has less corruption in the image to cause confusion). In addition, text fields on a catalog card are arranged in a coherent order which should make it possible to build a format-recognition system to facilitate the conversion of catalog cards to MARC records. Similar techniques, applied to full-text documents, will help identify the logical document structure from the layout structure of such documents.

**Table 1.1** Confusion Matrix (Test Set)

| | Actual | |
|---|---|---|
| **Predicted** | **Text** | **Noise** |
| Text | 1,015 | 14 |
| Noise | 26 | 33,160 |

## References

Breiman, L. et al. 1984. *Classification and Regression Trees*. Belmont, Calif: Wadsworth International Group.

Wong, K. Y., R. G. Casey, and F. M. Wahl. 1982. Document analysis system. *IBM Journal of Research and Development* 26:64"–56.

Project staff: John C. Handley, Associate Research Scientist, Charles T. Huff, Programmer/Analyst; Craig Henderson, Research Assistant

# Clustering Equivalent Bibliographic Records

**Project Manager:** Elaine Svenonius, Visiting Scholar

**Abstract**

This project assumes that information overload caused by Humphry Clinker type retrievals can be diminished by improving the method of presenting records to users. Techniques under investigation to improve record presentation include (1) grouping equivalent records into meaningful sets to reduce the number of records that must be scanned, and (2) providing a variety of ordered displays.

Retrieving too many records from a bibliographic database can cause "information overload." While normally associated with subject retrievals, information overload can also result from a search for a specific work. An example is the search for *Humphry Clinker* by Tobias Smollett. The OCLC author/title key *smol,hump* retrieves over 100 bibliographic records. Each record represents a version or edition of *Humphry Clinker*.

The *Humphry Clinker* problem (O'Neill and Vizine-Goetz n.d.), present in online computerized catalogs, is compounded in the large databases of bibliographic utilities and online regional networks. Such catalogs are global in nature, containing records for bibliographic items existing

in many collections. Searching a global catalog for a particular work is different from searching the catalog of a local collection because a local catalog is not likely to contain numerous bibliographic records representing the same work.

Standard approaches to reduce large retrieval sets such as qualification by subject, format or language are not appropriate or do not effectively reduce retrievals of the *Humphry Clinker* sort. Every record in the retrieval set represents the same work and therefore could be relevant to the user.

The point of departure in the present project is managing information overload by controlling how records are presented to the user. An ideal display suppresses information that is not useful and presents useful information in a helpful order. Such suppression can be achieved by clustering records that can be regarded as equivalent; helpful ordering can be achieved by offering users a variety of ways to scan retrieved items.

Clustering bibliographic records has four major tasks:

1. Define sets of equivalent bibliographic items
2. Test the reliability and validity of definitions
3. Determine whether the validated definitions can be used as a basis for algorithmic procedures to cluster and order bibliographic records
4. Develop displays of bibliographic records for different user interfaces

### Defining Equivalent Sets

The first task is to define sets of equivalent bibliographic items, where *item* means any physical manifestation of a work. To define these sets a variety of users' views must be encompassed that range from the casual reader, who regards as equivalent any legible copy of *Humphry Clinker*, to the sophisticated bibliographer interested in describing the publishing history of the work.

A tentative approach uses the following hierarchical ordering: work, text, typesetting, subedition, imprint, and reprint. These terms have entrenched and somewhat controversial meanings.

- **Work.** The set of all manifestations of an original text and all manifestations derived from that original by translation or revision. (This is somewhat narrower than some conventional definitions.)
- **Text.** The set of all items manifesting essentially the same content.
- **Typesetting (or edition).** The set of all items resulting from a single typographical composition.
- **Subedition.** The set of all items emanating from a single typesetting that exhibit publisher or publication variations.
- **Imprint.** The set of all items emanating from a single typesetting and published by one publisher, or a set of joint publishers, on one date.
- **Reprint.** The set of all subsequent printings of an item that preserves the image of a previous printing.

An example of a family-tree ordering of bibliographic items based on these definitions is given in figure 1.4. A linear ordering is obtained by listing all items from top to bottom in each branch, from left to right.

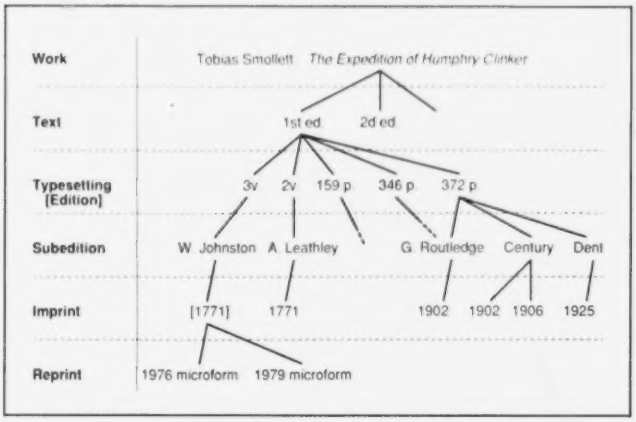



**Figure 1.4** Partial Family Tree for *Humphry Clinker*

A difficulty with these definitions is that they are conceptual in nature. To serve the purpose of clustering records in meaningful online displays, it is necessary to devise corresponding operational definitions. Operational definitions are constructive; in the present project they comprise sets of data elements occurring in bibliographic records. A *work* is the set of all items with the same author and title; a *text* is the set of all items with the same author, title, edition number, language, and translator; a *typesetting* is the set of all texts with the same number of pages; and an *imprint* is the set of all texts with the same publisher and publication date.

### Testing the Definitions

The second task is to test the operational definitions for reliability and validity. Is the definition *[author, title, edition number, language, translator]*, for example, necessary and sufficient to identify all manifestations of a text? The reliability and validity of the definitions will be measured in terms of precision (irrelevant items identified) and recall (relevant items missed). An example of a precision failure is an item incorrectly included in the set by virtue of variant cataloging or the accidental sharing of similar characteristics. An example of a recall failure is a text manifestation whose title differs from its mates.

The operational definitions will be tested by examining books obtained through interlibrary loan representing 200 English monographic works randomly sampled from the OCLC Online Union Catalog. The same sample set will be used to assess the extent of retrieval problems and the degree of set reduction that can be obtained by clustering records at the work, text, and typesetting levels.

### Algorithmic Clustering and Ordering Procedures

Implementing procedures for clustering records into equivalent sets depends on several factors. Chief among these is the feasibility of developing algorithms to recognize cataloging variations in the data elements that comprise the operational definitions. Using the 200 monograph sample, we will enumerate and categorize cataloging variations that occur for each data element, noting the degree to which variations are systematic. Such information will be useful in assessing whether the operational definitions can be used in automatic clustering procedures. Should the definitions prove useful, the next step is a long-term project to develop and test matching algorithms.

### Developing Displays

Many options are possible in online display, including suppressed equivalent records, user-defined sequences, and computer-user dialogues. The system could present bibliographic information, for instance, by displaying one entry for each different work. When the user identifies a work of interest, the system could prompt the user for a preferred edition, language, or sorting order (e.g., sorting by publisher and then by date or vice versa (figure 1.5). How to display bibliographic entries for multiply manifested works is part of a more comprehensive problem, namely, how generally to display records in large retrieval sets.

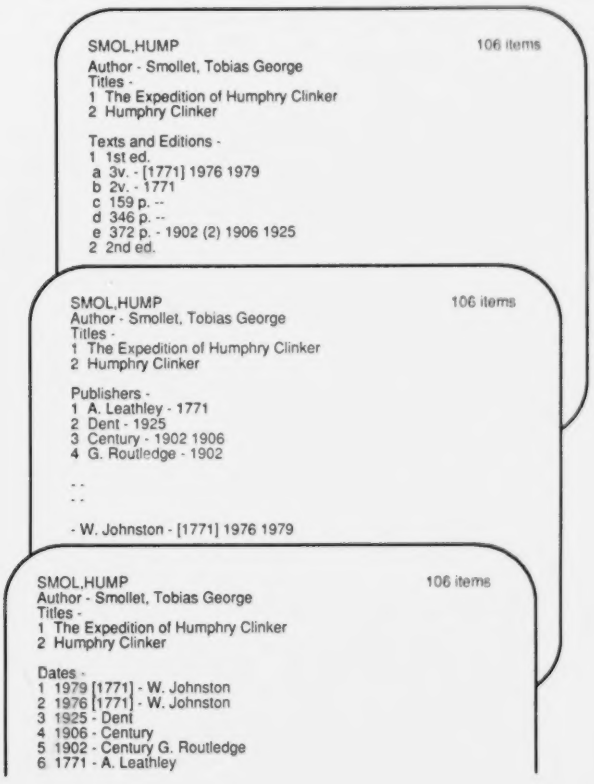



**Figure 1.5** Bibliographic Displays. Users could select displays ordered by texts and editions, publishers or dates.

### References

O'Neill, Edward T., and Diane Vizine-Goetz. n.d. Bibliographic relationships: Implications for the function of the catalog. In *The Conceptual Foundations of Descriptive Cataloging*, ed. Elaine Svenonius. Orlando, Fla.: Academic Press. In press.

Project staff: Edward T. O'Neill, Senior Research Scientist; Duane A. Rice, Research Assistant

## Collection Analysis

**Project Manager:** Edward T. O'Neill, Senior Research Scientist

**Abstract**

This study examines the monographic acquisitions of 11 Association of Research Libraries (ARL) members in two subject areas. Collections in botany and mathematical analysis from university libraries in the Committee on Institutional Cooperation (CIC, the Big Ten universities plus the University of Chicago) served as the study population. The OCLC and The Research Libraries Group (RLG) databases were used as the source of holdings information. The first phase, which examined the feasibility of automating collection analysis, found that the methodology provided a promising means of analyzing and comparing library collections. The second phase of the project will include development and testing of mathematical models of the collection development process.

The growing trend toward research library participation in cooperative collection development and resource sharing agreements has prompted librarians to seek consistent means to evaluate, compare, and analyze their collections. Unfortunately, most available methods are labor intensive. The purposes of this study are (1) to test the feasibility of using the databases of the bibliographic networks for computerized collection analysis, and (2) to develop and test mathematical models of the collection development process. The first phase is complete and the resulting paper is forthcoming in *College & Research Libraries* (Sanders, O'Neill, and Weibel n.d.).

The project was formally initiated in 1985 when OCLC and RLG were invited by CIC to participate in a meeting of its science bibliographers and collection development officers. The meeting explored the potential for cooperative collection development among CIC institutions. Discussions with participants and program planners suggested that science and technology collections would be good subject areas for the study.

### Sampling Methodology

Both objectives of the study required a sample of bibliographic records and related holdings information. Random samples of 500 monographic records from each of the two

subject areas, botany and mathematical analysis (which includes calculus, functional analysis, functions, and differential equations) were extracted from the OCLC Online Union Catalog. These two subject areas were selected because (1) their bibliographic character provides a useful contrast, (2) each was collected by all of the institutions, and (3) they were readily identifiable subjects in both the Library of Congress and the Dewey Decimal classifications. The samples were intended to be representative of recently published monographs in the subject areas, and thus in the pool of potential library acquisitions.

Only records with a copyright or publication date between 1978 and 1983 were included. This eliminated differential rate of retrospective conversion among the libraries as a factor in the comparison of holdings and minimized the effect of delayed acquisitions or cataloging backlogs. The records were then compared with related records in the Online Union Catalog to determine if any represented a publication with substantially the same content. Records for books with substantially the same content were excluded unless they represented the first book of the group entered into the database. Obvious articles from journals that had been cataloged separately and entered as monographs were also eliminated. After eliminating records not representing unique monographic texts, 392 records remained in the botany sample and 454 in the mathematical analysis sample.

As the sample was searched, all relevant OCLC holdings data were appended to the selected bibliographic records. However, because only 6 of the 11 CIC institutions are OCLC members, not all CIC member holdings were represented. Four of the institutions are RLG members, and one is not associated with any bibliographic network. To obtain holdings data for the RLG members, a listing of the bibliographic records in the samples was sent to RLG where it was checked against that database. To ensure the validity of the data, libraries were asked to verify their holdings.

## Findings

The concept of analyzing library collections by comparing current acquisition patterns with a pool of available monographs was found to be a valuable approach to collection evaluation. The acquisition patterns for both botany and mathematical analysis indicate that there is considerable potential for cooperative collection development among the CIC institutions. A relatively small shift in acquisition patterns could result in a significant reduction in the amount of material not acquired by any CIC institution. Such changes in collection development policies would reduce materials available to users locally but increase the availability of materials within CIC institutions. Whether the overall results of such changes would be desirable would depend on usage patterns, local expectations, and political conditions, none of which was examined in the study.

A high correlation between collection size and overlap was found. If this pattern is substantiated by further analysis, then libraries contemplating cooperative agreements might rely with some confidence on the more easily obtainable collection size statistics for a particular subject classification rather than computing common holdings.

From the analysis of the holdings it is clear that local library cataloging practices and bibliographic network policies affect the usability of online databases for collection analysis. Responses from CIC institutions indicate patterns of cataloging practices that require local validation to achieve reliability. Cataloging policies that resulted in partial cataloging of monographic series and no cataloging for some reserve, technical report, and theses collections were apparent.

The methodology proved to be a sound means of evaluating and comparing research collections. Potential uses for the results include accreditation reports, collection strength analysis for proposed new programs, cooperative project viability, and Conspectus validation. However, unless a method can be found to compensate for the unreported holdings, local validation of the holdings data is necessary to obtain consistent and reliable results.

## Modeling the Acquisition Process

The high correlation between collection size and overlap warrants investigation. Mathematical models of the collection acquisition process will be developed to explain and test several different theories of collection development. Particular aspects to be considered are the extent that local policies influence collections and the impact of random selection on the acquisition process. If valid models can be developed, they would add significantly to the understanding of multi-institution collection development and resource sharing.

## References

Sanders, Nancy P., Edward T. O'Neill, and Stuart L. Weibel. n.d. Automated collection analysis using the OCLC and RLG bibliographic databases. *College & Research Libraries*. Forthcoming.

Project staff: Nancy P. Sanders, Head, Home Economics Libraries, The Ohio State University, Columbus, Ohio; Stuart L. Weibel, Research Scientist, OCLC

# Collection Analysis System

**Project Manager:** David J. Stephens, Senior Programmer/Analyst

**Abstract**
A Collection Analysis system was developed using dBase III+ and compact discs. Libraries can use this system to compare their collections with selected peer groups; metric reports project the results of various collection development strategies. Collection developers can view lists of specific bibliographic items not held by their institutions but held by specified numbers of peers within a selected peer group.

The OCLC Collection Analysis system has evolved into a prototype that utilizes a microcomputer workstation, compact disc-read only memory (CD-ROM), dBase III, and a subset of the OCLC Online Union Catalog. The new prototype provides a flexible user interface, metric reports, and a large database stored on compact disc.

## Prototype Features

Users may select peer groups like those shown in figure 1.6. The comparative size value of 87% for the medium ARL peer group means that the evaluating library holds 13% fewer titles than the average peer in this group. The average holdings value gives the average number of peer group members who hold an item in terms of a percentage of that peer group's members. Thus, on average, items in the medium ARL peer group are held by 24.6% of its members.

| Peer Group | Comparative Size (%) | Average Holdings (%) |
|---|---|---|
| All Academic Research Libraries in OCLC | 105 | 15.2 |
| Large ARL (Largest 25%) | 71 | 26.4 |
| Medium ARL (2nd Largest 25%) | 87 | 24.6 |
| Small ARL (3rd Largest 25%) | 123 | 16.1 |
| ARL Private Institutions | 134 | 19.5 |
| Competitive Colleges | 65 | 16.2 |

**Figure 1.6** Peer Group Selection

Once a peer group is chosen, metric reports can be generated and used to compare subject areas within the evaluating institution's collection. This allows the institution to identify subject areas needing greatest attention. A sample report is shown in figure 1.7.

The subject area analysis report gives the status of the evaluating institution relative to the peer group for 500 broad, predefined subject areas. It shows the current number of titles held by the evaluating institution and the average peer in each subject area. It also shows the effectiveness of a most-held-first acquisition strategy whereby titles held by the greatest number of peers are acquired first.

Title frequency within the peer group is measured by the selection threshold. With a 50% selection threshold an item must be held by at least half the peer group's members before it is included in the selected titles column. The comparative size gives the quotient of expected number of titles (the sum of current titles plus titles above the threshold) divided by number of titles held by the average peer. This value can be used to judge the appropriate acquisition effort for each subject area. If this number reaches 100% at a high selection threshold such as 50%, the evaluating institution has not followed a most-held-first strategy in the past.

Peer Group: Medium ARL (2nd Largest 25%)

| Subject Area | Selection Threshold (%) | Selected Titles | Comparative Size (%) |
|---|---|---|---|
| QA300–QA433 | | | |
| Mathematical Analysis | 70 | 3 | 92 |
| | 50 | 5 | 108 |
| Current Evaluator Titles = 12 | 30 | 7 | 125 |
| Current Comparative Size = 66% | 10 | 10 | 150 |
| | 0 | 25 | 275 |
| QA440–QA799 | | | |
| Geometry/Trigonometry | 70 | 5 | 120 |
| | 50 | 25 | 160 |
| Current Evaluator Titles = 55 | 30 | 51 | 212 |
| Current Comparative Size = 110% | 10 | 93 | 296 |
| | 0 | 186 | 482 |
| QA801–QA939 | | | |
| Analytic Mechanics | 70 | 3 | 65 |
| | 50 | 7 | 71 |
| Current Evaluator Titles = 42 | 30 | 12 | 77 |
| Current Comparative Size = 54% | 10 | 22 | 90 |
| | 0 | 34 | 106 |

**Figure 1.7** Subject Area Analysis Report

In the sample shown in figure 1.7 the evaluating library holds 12 titles, 66% of the number held by a typical peer in the mathematical analysis subject area. There are five titles held by 50% or more peers. If these titles were acquired, the evaluating institution would hold 108% of titles held by the average peer. Figures for the geometry/trigonometry subject area reveal that a most-held-first acquisition strategy was not followed; the analytical mechanics subject area shows results of such a strategy. Once collection metrics have identified subject areas needing attention, the bibliographic comparison can be used to show specific items not held by the evaluating institution but held by peer group members at a specified selection threshold (figure 1.8).

More information about items listed in the bibliographic comparison are found using a zoom facility. The collection analyst may want to consider whether to acquire the items listed. Of course, an evaluating institution may proceed directly to bibliographic reports of subject areas of interest without using collection metrics.

Bibliographic Comparison
Peer Group: Academic Research Libraries in OCLC

| LC Class | Title Statement | Peer Group Holdings |
|---|---|---|
| QA300 | A source book in classical analysis | 49 |
| QA300 | Lectures on analysis | 46 |
| QA300 | Selective computation | 38 |
| QA300 | SIAM journal on mathematical analysis | 60 |
| QA300 | Undergraduate analysis | 38 |
| QA300.5 | Seminar on micro local analysis | 74 |

**Figure 1.8** Title Selection

### Field Test

OCLC conducted a field test in December 1987 at four test sites. The purpose of this test was to gauge interest for such a tool in the library community and to better identify needed functionality.

Response by the library community indicates that the ability to compare a library with groups of other libraries is valuable. Such comparisons may enable libraries to

1. direct collection budgeting across all major subject areas of the collection;
2. aid departmental budgeting;
3. assist in the selection of titles for acquisition;
4. assist librarians working with scholars and faculty to evaluate specialized areas of a collection;
5. assist the accreditation process by allowing comparisons with groups of peer libraries;
6. project acquisition costs for new programs;
7. evaluate collection development policies;
8. assist the cooperative development of collections among groups of libraries.

Field test results made apparent the need to refine criteria used to select records for the compact disc. In particular, duplicate records need to be reduced and serials either need to be eliminated or separated from monographs. Also, date cutoffs and foreign language issues need to be investigated. The field test also indicated that further investigation into collection metrics is required.

### Future Research

Future research will focus on (1) developing better collection metrics, (2) finding better selection criteria for bibliographic records to be stored on compact disc, and (3) developing a cost effective means of allowing institutions to define peer groups.

Project staff: Martin Dillon, Director, Office of Research; Mark Crook, Consulting Systems Analyst; Kevin Flash, Systems Analyst, Development Division

# Connectionist Models of Computation: Their Potential in Information Systems

**Project Manager:** Stuart L. Weibel, Research Scientist

**Abstract**
Connectionist models of computation represent a novel approach to solving problems of complex pattern recognition in the information science field. Areas of potential contribution include document analysis, optical character recognition, and information retrieval. Our exploratory efforts are intended to help us monitor the potential for progress and anticipate the contributions that this new research area may provide.

Real-time pattern recognition remains one of the more difficult computer processing tasks. Problems of complex pattern recognition are addressed by computer science theories and methods known variously as "connectionism," "neural nets," or "parallel distributed processing." These approaches can be applied to pattern recognition problems encountered in library and information science.

### Connectionist Models

The essential model for neural net processing is based on many simple processing units (corresponding to biological neurons) associated in richly interconnected networks. The power of this model derives from the connections among these units, just as neural power is thought to reside in the complexity of neural interconnections.

Processing units, or nodes, have inputs and outputs, just as neurons have dendrites and axons. A unit typically combines the inputs according to an "activation rule" and applies the result to its output connections, which in turn feed the inputs of other units. The strength or weight of a connection varies and is determined by a learning rule that specifies how connection weights should change in response to changing input. Thus, if a given set of connection weights brings the network closer to the desired output, these weights are reinforced. In this manner, a network will converge toward a correct answer and stabilize during a training period. A stable network implies that the desired recognition knowledge is encoded in the activation state of nodes and the pattern of connection weights of the network.

Network nodes fall into three classes (input nodes, output nodes, and hidden units) arranged in layers, as illustrated in figure 1.9. For a character recognition network, the input layer might have a binary node for each pixel and eight binary output nodes to represent an extended ASCII code. The number, size, and character of the hidden layers is problem-dependent and empirically determined.

Connectionist models of computation are presently a speculative area of research with a controversial history.

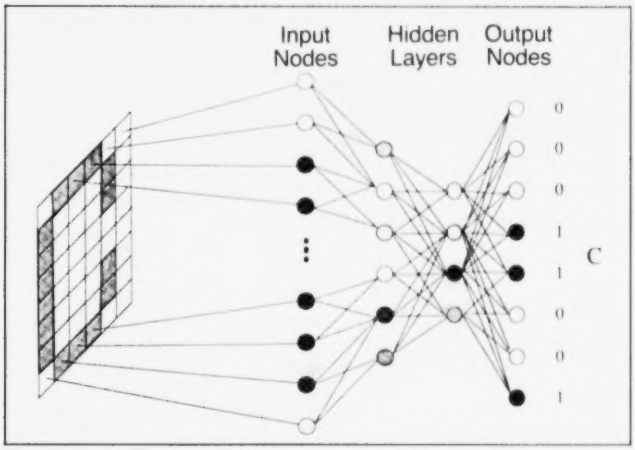



**Figure 1.9** Network Nodes. Hypothetical network configuration for character recognition.

The field was born in the 1960s amid new optimism about the power of computers and understanding biological nervous systems, but subsequently fell into decline. Recent advances in computing speed and parallel hardware and a better theoretical foundation have rekindled interest. Several factors make these models appealing.

**Subsymbolic knowledge representations.** Expert systems have become popular for capturing narrow expertise about a constrained domain. This kind of high-level knowledge often lends itself to the formulation of rules, but human cognition is not easily embodied in a list of rules in either computer or natural languages. It has been demonstrated that this kind of subsymbolic knowledge is readily encoded in richly connected neural networks.

**Learning.** The ability of connectionist systems to learn is certainly a major factor in their rising popularity. Many demonstrations of this capability have established this aspect of neural nets as an important characteristic that is particular to computer science. The design of networks and implementation of efficient and effective learning algorithms are currently major topics of research, but there is no question that such systems can learn significant pattern recognition and classification tasks.

**Parallelism.** One of the driving forces in the movement toward connectionism is increasing interest in parallel processing architectures. The availability of parallel processing machines, still largely experimental, will require new algorithms that depart from conventional serial approaches to computation. Connectionist models are intrinsically parallel in nature and may represent a natural way to take advantage of parallel architectures.

**Tolerance for ambiguous, noisy, and incomplete data.** Human cognition often takes place with scant input data—one can pick out familiar objects in a crowd or out of the corner of one's eye. Systems that perform similar kinds of pattern matching will find wide applicability.

## Connectionist Models at OCLC

We are developing an array of neural net simulation tools to experiment with these techniques in information systems research. Our near-term goal is a tool kit that will enable the rapid specification of networks using a graphical mouse-driven interface. We will use these tools to explore problems relevant to ongoing work that may benefit from the connectionist approach to pattern recognition.

An example of such a problem is the improvement of optical character recognition (OCR) performance. Our preliminary experiments demonstrate that a relatively simple network can learn to recognize characters in the IBM PC® character set. Testing a wide range of character fonts will enhance our understanding of the problems associated with improving the performance of existing OCR technology.

We also expect to use connectionist models to identify and classify layout objects in documents. These are important aspects of automating document structure analysis, which is used to build databases from paper document collections.

Information retrieval represents a further opportunity for application of neural networks. Retrieval is predominantly a pattern matching activity; queries must be represented as a pattern of information and matched against the contents of a database. It is conceivable that connectionist systems will assist this intrinsically imperfect task in future information systems.

## Conclusion

The success or failure of connectionist paradigms hinges on the ability of these techniques to solve problems such as pattern recognition. Our exploratory efforts are intended to help us monitor the potential for progress and anticipate the contributions of this new field.

---

Project staff: Craig Henderson, Research Assistant; Charles T. Huff, Programmer/Analyst

# Database Quality

**Project Manager:** Edward T. O'Neill, Senior Research Scientist

**Abstract**
This ongoing project investigates various issues relating to the quality and usability of the OCLC Online Union Catalog. Current research activities include (1) assessment of data quality, (2) studies of automatic error detection and correction algorithms, and (3) design and operation of authority control systems. Much of the current research is undertaken to support the Subject Heading Correction project.

The goal of database quality activities is to improve the quality and usability of the OCLC Online Union Catalog. Project staff conduct research and support the Subject Heading Correction project.

## Subject Heading Correction Project

The Subject Heading Correction project is developing an OCLC database of unique subject headings and correcting subject headings in the Online Union Catalog. When complete, the OCLC subject database will have many of the characteristics associated with an authority file. The scope of the project is limited to Library of Congress (LC) topical and geographic subject headings (MARC fields 650 and 651). The project has three major steps: (1) develop an initial OCLC subject database containing all Library of Congress subject headings (LCSH), (2) identify and correct erroneous subject headings, and (3) correct bibliographic records in the Online Union Catalog.

**Initial database development.** The *Library of Congress Subject Headings (LCSH)* is the basis for the OCLC subject database. The database contains all established headings and cross-references from LC's subject authority and 19.7 million subject headings from OCLC bibliographic records. The OCLC control number links subject headings to bibliographic records. The OCLC subject database contains over four million unique subject headings; fewer than 4% come directly from *LCSH*.

**Correcting the subject database.** The initial quality of entries in the OCLC subject database varies widely. Entries from LC subject authority records contain few errors. Entries from LC bibliographic records have a higher, but still relatively low, error rate while headings found only in contributed records have the highest error rate.

The approach to error correction uses automatic detection of erroneous headings and, when feasible, the identification of the correct heading. The error correction activity will identify and correct four major types of errors found within subfields: (1) variation in the form of subfield text, (2) obsolete subfields, (3) incorrect MARC tag or subfield code, and (4) typographical and minor spelling errors. The correction procedure includes the following steps:

1. Selecting preferred form of subfield texts
2. Correcting subfield code errors
3. Identification of invalid subfields
4. Correcting tagging errors
5. Correcting typographical errors
6. Linking manually identified variant subfields
7. Identifying uncorrectable subfields

All correction is done by linking the erroneous or variant entries to their valid, current, or preferred form. All original information and linkage to the bibliographic records are retained through the error correction process. As a result, any correction made to the subject database is fully reversible. This reversibility is an essential element in the correction strategy. No matter how good a computerized algorithm is or how careful the manual review, some valid entries will be changed.

**Correcting the Online Union Catalog.** After the error correction phase, the bibliographic records themselves will be corrected. Since the subject database is linked to the bibliographic records by the OCLC control number, every bibliographic record with an erroneous subject heading can be identified and corrected. Before changing a subject heading, the bibliographic record will be checked to see that the erroneous form of the heading remains, unmodified, in the record. Subject headings that have been changed since the subject database was generated will not be revised.

## Research Activities

The current Subject Heading Correction project will correct many, but not all, of the errors in LC subject headings. Several major types of errors that will not be systematically corrected include compound subfields, abbreviation variants, qualification errors, inversion errors, invalid geographic subdivisions, the use of cross-references, and subdivision assignment inconsistent with subject authorities. It is known that rarely used headings are frequently incorrect and the analysis of these headings would identify a wide variety of other types of subject heading errors.

**Correction algorithms.** Research is being planned to identify and correct common errors that are not being considered in the initial phase of the Subject Heading Correction project. Algorithms are being developed to identify and/or correct most of the errors remaining after the initial correction of the subject database. These procedures cannot be fully tested and evaluated until the more common errors in the subject database are corrected.

**Spelling correction.** Research on spelling correction is becoming a significant part of the Database Quality project. As a class of errors, "misspelling" is a generic term that includes spelling as well as typographical, scanning, transmission, and other similar types of errors. Since

typographical errors are the largest class of error in the Online Union Catalog, spelling correction algorithms will be an important aspect of any database quality control project.

A major review and analysis of published research on quality of online databases and the means of controlling database quality was completed. The primary result of this study is a chapter for the 1988 *Annual Review of Information Science and Technology* on "Quality Control in Online Databases" (O'Neill and Vizine-Goetz 1988). In addition, the project team reviewed the problems of database quality control in general. Some of the methodologies used by other database producers may be adaptable for use at OCLC.

**Authority control.** A continuing area of research is authority control systems. The size of the OCLC database, combined with the number and diversity of users, suggests that, in the long term, an interactive authority control system with automatic error detection and correction will be required. The cataloger entering records is in the best position to know which of the proposed corrections are appropriate or if the entry is, in fact, valid. It is more difficult for OCLC staff to attempt to correct records without access to the materials being cataloged. Even when the cataloger cannot validate the records, the entering library can save records for further review.

Authority control research at OCLC has focused on developing advanced error detection algorithms for use with authority files to overcome the rigidity normally associated with authority control. The system should provide the cataloger with as much information as possible rather than respond with *Invalid Entry* or other similar terse message. When a bibliographic entry is made that does not match an authority record, for example, the error detection algorithm would be invoked. The algorithm would produce a short list of potential corrections. For subject headings such a list typically would have one or two headings and, in most cases, it would contain the valid form of the heading. The user could then select the valid heading and correct the bibliographic record.

### References

O'Neill, Edward T., and Diane Vizine-Goetz. n.d. Quality control in online databases. *Annual Review of Information Science and Technology 1988*. Forthcoming.

Project staff: Karen J. Rankey, Senior Programmer/Analyst; Diane Vizine-Goetz, Research Scientist

## Document Facsimile Transmission Studies

**Project Manager:** John C. Handley, Associate Research Scientist

**Abstract**
This project studies the feasibility of document delivery through image transmission. The technical aspects of International Telegraph and Telephone Consultative Committee (CCITT) group 3 and group 4 standards for image compression and digital transmission are considered. A facsimile workstation is described and a model of a potential system for document delivery is outlined.

This project seeks to determine the feasibility of electronic document delivery through facsimile transmission. This requires technology assessment and models of document delivery systems such as interlibrary loan in which facsimiles could play a part.

### Standards

A document facsimile is an encoded black-and-white image of a document. International standards for facsimile equipment were developed under the direction of the International Telegraph and Telephone Consultative Committee (CCITT) to specify the characteristics of the transmission and encoding of documents. The current widely used standard is for group 3 equipment. It specifies that an International Standards Organization (ISO) A4 (8.27" × 11.69") document page be transmitted in a digital format over general switched telephone lines at a minimum of 2,400 bits per second (bps) in about one minute.

The compression scheme to encode the digitized image was optimized using run-length statistics for English typewritten pages, handwriting, documents in several languages, and line drawings. The standard assumes a lower quality transmission medium (public switched telephone network) and thus allows for a few errors in transmission. The maximum width is 1,728 pixels in 8.46 inches. The maximum height is 2,432 pixels in 11.9 inches. This limits the horizontal and vertical resolutions to about 204 dots per inch (dpi). This resolution is adequate for office documents such as letters and memos but is insufficient for typeset documents, especially technical ones. Therefore, the group 3 standard is not adequate for a document delivery system consisting of journal or magazine articles.

A group 4 standard was established assuming error-free transmission over public data networks, packet-switched networks (X.25), integrated services digital networks (ISDN), or public switched telephone lines (with the appropriate equipment). In addition to allowing images to be transmitted uncompressed, the standard uses a two-dimensional run-length encoding scheme similar to group 3 encoding options. It allows image resolutions of 200, 240, 300, or 400 dpi, both horizontally and vertically. At 300

dpi, most documents are reasonably legible. At 400 dpi, the quality is similar to a photocopy, suitable for documents containing text and line drawings. Digitized halftone images pose a special problem. They are of generally low quality when obtained from a bilevel scanner and do not compress well with run-length encodings such as the type used in the group 4 standard.

At 300 dpi, an encoded digitized document page as large as 600,000 bytes is not uncommon, especially if it contains halftone images such a photographs or artwork. At 9,600 bps it would take over eight minutes to transmit such a page. A typical eight-page journal article would take about 30 minutes to transmit; at 56,000 bps, just over 5 minutes. It is clear that transmission over 9,600 bps telephone lines is unsatisfactory. Data transmission rates of 56,000 bps (packet switched) or even 64,000 bps (ISDN) are preferred.

### Facsimile Workstation

A facsimile delivery workstation consists of a personal computer with a network interface, a laser printer, and optionally a large, high-resolution display monitor. The workstation would receive a file containing the compressed digitized image of a document and store it on disk or in main memory. A user would then either print the document on the laser printer or display it. To print or display the document, it must be decompressed or decoded. Currently there is no hardware commonly available to do the decompression so it must be done by software. Software decompression is somewhat slower but is more flexible and can take advantage of faster general purpose central processing units (CPUs). To capture an image for transmission, an optical scanner would be used.

### Document Delivery System

One model of a document delivery system would store documents as they are requested. Subsequent requests are filled by documents from a cache, saving the costs of retrieving the hard copy, scanning, and transmission. To analyze this scenario, we need to model the arrival of duplicate requests. Efron and Thisted (1975) give guidance for such a model. A straightforward stochastic model was derived for the rate of arrival of duplicate requests. One can thus estimate and forecast total costs and revenues. Preliminary data based on a week's worth of interlibrary loan requests for journal articles show that the rate of duplicate requests is small compared to the overall request rate. It is generally believed that the duplicate request rate would rise due to the growth of the document cache. The stochastic model allows for the rates to change so that once the behavior is known it can be used in the model.

Document delivery through image transmission will soon become cost effective. Having robust and flexible models of potential systems will allow us to maximally exploit the technology.

### References

Efron, Bradley, and Ronald Thisted. 1975. *Estimating the number of unseen species (how many words did Shakespeare know?)* Technical Report no. 9 (July 25). Stanford, Calif.: Stanford University, Division of Biostatistics.

# Duplicate Detection and the "Species Problem"

**Project Manager:** John A. Bunge, Associate Research Scientist

**Abstract**

This project seeks to estimate the number of unique records in the OCLC Online Union Catalog using automated statistical techniques. The problem of duplicate detection is described in statistical literature as the "species problem"; duplicate records in a database are members of the same species. Computer simulations are used to investigate the complex properties of the statistical procedures. A robust estimator based on the samples of entire species performs exceptionally well.

Bibliographic records in the OCLC Online Union Catalog can be grouped or classified in many different ways. Records could be grouped into "duplicate classes," with each class containing some number of duplicate records. For example, there might be a class consisting of $n$ records for the same edition and printing of Mark Twain's *Huckleberry Finn*. The total number of classes would be the true number of distinct records in the database. In other words, if there are $n$ records in the database (currently $n$ = about 18 million) and these records can be grouped into $k$ duplicate classes, we would like to have a database with exactly one representative of each of the $k$ classes, i.e., a database containing only $k$, not $n$, records. However, $k$ is unknown.

Online Union Catalog records could be grouped using other criteria such as derived search keys. For example, there might be nine records with the title key *AAA,BB,CC,D*; one with the key *AAA,BB,CC,E*; and so on. Again, this method groups all $n$ records into $k$ classes (here $k$ is the number of title keys, currently about 5.8 million).

From a mathematical point of view, these two grouping methods are different versions of the same problem. In each instance a population of $n$ individuals (bibliographic records) is divided into $k$ classes. Each class contains a certain number of individuals (1, 2, 3, . . . $n$) and there is an unknown number of classes with only one record. The number of classes with one member would be the number of records occurring only once in the database. If we knew this number, we would know how many records should be examined for potential duplicates.

Ideally, we would run a program that would first group the records in the desired way and then count the records in each class and the number of classes. The Online Union Catalog, however, is so large that it is not often feasible to run such a program on the whole file. Instead, we must draw a sample of database records. We then can group the sample records and count the class sizes and number of classes. The question is, how can we use the sample information to tell us something about the population? In the duplicate detection context, how can we use the counts of unique and duplicate records in the sample to tell us something about the numbers of unique and duplicate records in the Online Union Catalog? If a record is unique, what can we say about this item? We cannot conclude that it is unique in the population because, even though it only appeared once in the sample, it may actually come from a class of size two, three or more in the population. In fact, many items that are unique in the sample are not unique in the population—it just happened that only one member of the class made it into the sample. Similarly, a class of size two in the sample might actually be a class of size two in the population, but it could also be two items from a class of size three, four or more in the population.

Statisticians have studied this general problem under the name "the Species Problem" because another version of the problem arose in entomological studies. From a random selection of butterflies, for example, some species will be represented many times, some several times, and some only once. From this information, how can we estimate the number of species not collected? What can be said about the species that are not represented in the sample?

The basic difficulty in the species problem is that we are sampling the wrong thing: we should be taking a random sample of whole species or classes but instead we are taking a sample of the individuals or items. In the duplicate detection context, suppose we wish to estimate the average size of the duplicate classes. We ought to take a random sample of the duplicate classes such as *Huckleberry Finn* and include all *n* members of its class in the sample.

There is not a great deal of statistical literature relevant to this problem; however, some techniques have been developed. Of these techniques, some perform so poorly they are useless, but others offer some promise. Using computer simulation, we are currently studying the performance of these statistical methods. A software workbench of UNIX™ utilities was written consisting of programs that generate populations, make random samples, and compute statistics. The statistical procedures may be investigated by using these programs and then comparing the results with the true population. If the estimate is consistently close to the true population (unbiased with a small variance) over many large and different samples, it can be expected to be accurate when used on data for which the true population value is unknown such as the number of unique records in the Online Union Catalog.

To emulate a database of bibliographic records and have a meaningful sample, we must generate natural populations, a difficult task due to the wide variety of possible populations. Two methods were developed: (1) the deterministic method which is "unnatural" in the sense that every class size up to the maximum is represented, and (2) the truncated geometric method which is random and "natural" in the sense that the population is generated using random numbers and thus some class sizes do not appear in the sample.

The software allows four ways of random sampling. We can sample individual records or an entire class. In each case we can sample with or without replacement. By storing a minimum amount of information, random samples are quickly obtained from arbitrarily large populations. This enables us to take many samples to test fully the performance of the statistics on a population as large as the Online Union Catalog.

Two types of statistics are used to describe the species problem: (1) statistics based on samples of individuals without replacement, and (2) statistics based on samples of entire classes without replacement. Our simulations reveal that the unique unbiased estimator of the number of classes based on samples of individuals gives values that vary widely. Its behavior is far too aberrant to be used in general but there are situations when the variance of the statistic is small enough to be useful. Sampling individuals is much easier to do on real databases than sampling classes.

The statistic based on class sizes performed remarkably well. The statistical method we used is an approximation to a statistic that will theoretically always work, but is computationally intractable (Andreatta and Kaufman 1986). Based on preliminary simulations of population sizes of 1,000,000 and sample sizes of 100, the standard deviation of the estimated number of unique items is less than 5% of the true population value. The Andreatta-Kaufman statistic can be expected to perform better for larger sample sizes and large populations.

## References

Andreatta, G., and G. M. Kaufman. 1986. Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association* 81:657–66.

Project staff: Mark Crook, Consulting Systems Analyst; John C. Handley, Associate Research Scientist.

# Graph-Text

**Project Manager:** Thomas B. Hickey, Consulting Research Scientist

**Abstract**

The Graph-Text project put one volume of a chemical encyclopedia on compact disc. The prototype system uses a modified version of the OCLC Search CD450™ system user interface and retrieval system. Users can search the text and display or print text with embedded graphics such as figures and tables. Links allow users to jump between text, figures, and tables, and articles are enhanced with automatically generated tables of contents.

Graph-Text is a system for the creation, retrieval, and display of documents containing text, graphics, equations, and tables. Databases are created from existing machine-readable files or using optical character recognition (OCR) scanning and displayed on bit-mapped display and printing devices.

During the past year, we put one volume of the 25-volume *Kirk-Othmer Encyclopedia of Chemical Technology*, including graphics, on compact disc-read only memory (CD-ROM) for demonstration. This volume occupies approximately one-thirtieth of the available compact disc (CD) space. Each article of the encyclopedia is a directory on the CD containing the file for printing/display and the graphics associated with the article. The files are stored on the CD using the High-Sierra file format.

In addition to this database conversion, printing and display systems were rewritten and merged using the same software for both functions. This simplified the system, ensured consistent results, and speeded printing to acceptable rates. To support multiple display environments and enable improvements to the user interface, a simple portable window system was written that supports several displays under different operating systems. The system is currently running on IBM PCs supporting EGA, Hercules®, and WY-700® display modes, as well as on Apollo® and Sun® Workstations. Mouse control was added for workstations that have pointing devices. Figure 1.10 shows the system running on an Apollo computer.

The structure of the database was parsed and automatically tagged using the Standard Generalized Markup Lan-

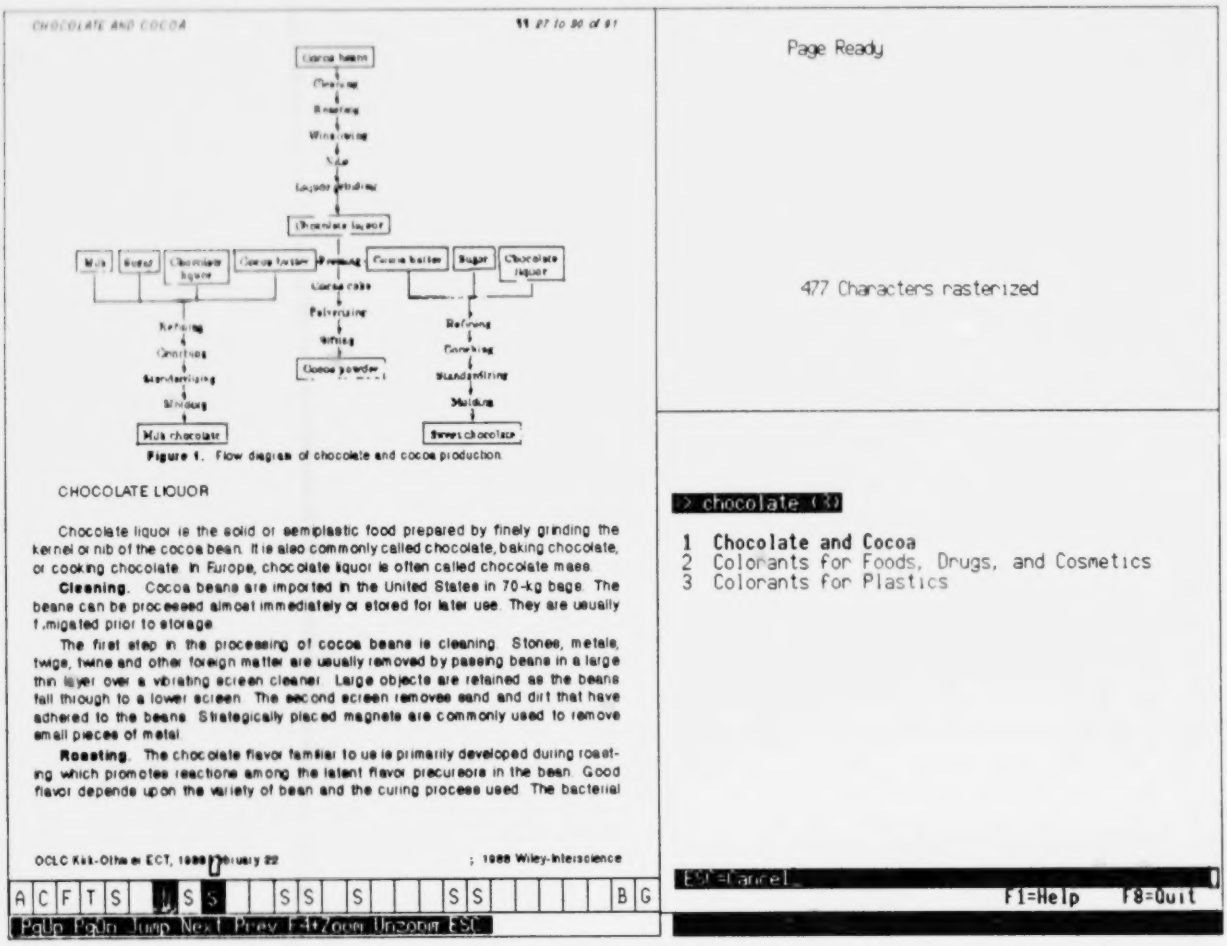

CHOCOLATE AND COCOA
Figure 1. Flow diagram of chocolate and cocoa production.
CHOCOLATE LIQUOR
Chocolate liquor is the solid or semiplastic food prepared by finely grinding the kernel or nib of the cocoa bean. It is also commonly called chocolate, baking chocolate, or cooking chocolate. In Europe, chocolate liquor is often called chocolate mass.
Cleaning. Cocoa beans are imported in the United States in 70-kg bags. The beans can be processed almost immediately or stored for later use. They are usually fumigated prior to storage.
The first step in the processing of cocoa beans is cleaning. Stones, metals, twigs, twine and other foreign matter are usually removed by passing beans in a large thin layer over a vibrating screen cleaner. Large objects are retained as the beans fall through to a lower screen. The second screen removes sand and dirt that have adhered to the beans. Strategically placed magnets are commonly used to remove small pieces of metal.
Roasting. The chocolate flavor familiar to us is primarily developed during roasting which promotes reactions among the latent flavor precursors in the bean. Good flavor depends upon the variety of bean and the curing process used. The bacterial
OCLC Kirk-Othmer ECT, 1988
; 1988 Wiley-Interscience
PgUp PgDn Jump Next Prev F4:Zoom Unzoom ESC
Page Ready
477 Characters rasterized
> chocolate (3)
1 Chocolate and Cocoa
2 Colorants for Foods, Drugs, and Cosmetics
3 Colorants for Plastics
ESC=Cancel
F1=Help F8=Quit

**Figure 1.10** Graph-Text Display

guage (SGML). This tagged text was then used for indexing and subsequent conversion into the TEX™ typesetting language. The SGML tags are carried throughout the system. Links are available between the text, figures, and tables, and graphics are displayed embedded in the text.

In addition to searching using a modified version of the OCLC Search CD450 system full-text retrieval interface, users can browse selected articles, print pages or whole articles on a laser or impact printer, and zoom in on text and graphics.

The Graph-Text project is in its final phase as a research project. Work is continuing to document the project, and the techniques developed are being transferred to development.

Project staff: Jeffrey L. Gerckens, Research Assistant; Sehchang Hah, Research Assistant; John C. Handley, Associate Research Scientist; Meichin Chung, Programmer

## Information Display

**Project Manager:** Thomas B. Hickey, Consulting Research Scientist

**Abstract**
This project develops and tests abstractions related to the display of text and graphics on computer monitors. The project builds on knowledge of user needs and desires gained during the Graph-Text project and experience in developing applications for those needs. It concentrates on the display of text and graphics on screens with gray-level shading in a manner that has the flexibility to accommodate individual needs and preferences.

Research in information display stems from several trends and assumptions:

1. Much of the reading that is currently done from paper will gradually shift to terminals over the next decade.
2. 100 dpi color/gray scale displays will become the norm (3-5 years).
3. A large amount of processing power at the workstation will soon be available (three times the power of current 68020 and 80386 processing chips).
4. There are several parameters such as layout and fonts for which it is possible and desirable to provide user control.

The project will proceed in three overlapping stages:

1. Identify parameters such as font size, shading, and layout that should be adjustable by the user. Experience, a review of literature, and experimentation will help determine these parameters.
2. Develop tools that allow users to manipulate parameters. Tools will be based on PostScript® and Metafont™ imaging systems, use the X Window System for display, and support display of images and graphics using scaling and shading.
3. Incorporate tools into document display systems for experimentation. A schematic of such a system showing the software relationships among the user interface, retrieval system, and window system appears in figure 1.11.

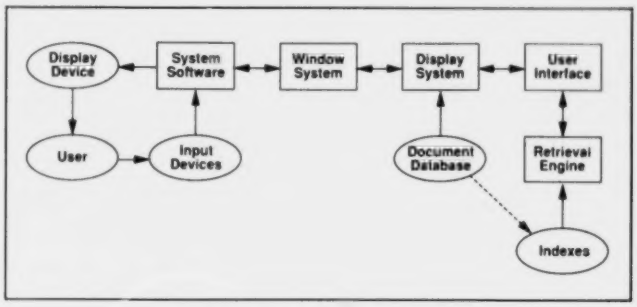



**Figure 1.11** Software Schematic

The Information Display project is still in the planning stages, but some programming work has been done. A simple PostScript interpreter written in the Ada programming language and using the X Window System for display has been written, as have some experimental scaling and gray-scale programs.

Project staff: Thomas L. Terrall, Senior Programmer/Analyst

## Information Services Project

**Project Manager:** Mark Crook, Consulting Systems Analyst

**Abstract**
The Information Services (IS) project uses statistical and data processing tools to answer questions about OCLC's databases. Questions are submitted by OCLC staff and scholars outside OCLC. This year efforts have focused on four classes of questions: descriptive statistics, marketing-related data, database-management support, and product-development support. Plans for future directions are presented.

Since its inception in 1987, the Information Services (IS) project staff have answered over 100 questions about OCLC databases. The examples for each type of question given here represent an array of the questions that IS han-

dled from January through June 1988. Answers were supported by appropriate statistical methods and data processing practices.

**Descriptive statistics.** These questions represent the majority of those asked. They often deal with frequencies of record types and the contents of the records. A descriptive statistical response can either answer the question or establish the scope for another question. Examples of such questions that have been asked include:

- How many festschriften are there in the OCLC Online Union Catalog?
- What is the frequency distribution of LC classes across the OCLC Union Catalog?
- How many pre-1800 imprints are there in the OCLC Online Union Catalog?
- How many records have a given publisher?
- How many records have a given number of holdings?
- Who are the most popular authors?

**Marketing-related data.** Questions in this category often ask for a comparison of the OCLC Online Union Catalog with another database. These answers are more complex than those based on descriptive statistics in that they are based on the data derived from both databases. Answers generally take more time to generate and require more care in setup and methodology. Examples of such questions:

- What amount of overlap is there between a given database and the OCLC Online Union Catalog or a subset of the OCLC Online Union Catalog?
- How well does the CAT CD450 database meet the needs of the target user?

**Database management support.** These questions deal with issues of size, growth, and performance of the OCLC Online Union Catalog, related systems, and subset databases. IS uses third-party software in some cases to discern the answers. The answers help requestors make informed decisions about system hardware purchases, marketing efforts, and improving system performance. Examples of such questions:

- What is the level of performance of the current system configuration in relation to its capacity?
- What is the size, growth rate, and complexity of the Chinese, Japanese, and Korean (CJK) subset database within the OCLC Online Union Catalog?
- Questions dealing with the current tapeload program algorithm's performance and with duplicate detection.
- How have the California State Project participants been using the online system?

**Product-development support.** These questions are the most complex and require advanced statistical methods and modeling techniques. IS provides data and interpretive, inferential insight. Thus, product managers and developers can make decisions and projections about products. Examples of such questions:

- Given certain scope parameters, what would the economies of a FAX-based Interlibrary Loan system be?
- From a sample, what can we extrapolate about a given subcollection from the OCLC Online Union Catalog, e.g., music, Latvian records, subject-specific subsets, etc.?
- What is the best record selection strategy for a CD-based system such as CAT CD450, CD-Resource Sharing, or Search CD450?
- How could we reduce large retrievals and enhance small ones on the new reference system?
- What is the nature of a browsable title index given certain criteria?

IS has a collection of software tools to manipulate records of all types. Most of the software is written in the Statistical Analysis System® (SAS) fourth generation language (4GL) and can be easily adapted to answer slightly different questions. The use of SAS and dBase programs permits custom applications to be generated in hours instead of days. Library routines permit IS to reduce the amount of turnaround time still further. IS has the powerful statistical procedures of SAS in both the mainframe and personal computer environment, dBase on the personal computer and Macintosh, and the Classification and Regression on Trees package (CART) on the Sun® Network.

### Future Plans

IS is expanding its services to the scholarly community outside OCLC through a pilot program called Scholarly Support (SS). Its goal is to derive hidden information from the OCLC database to support research in other disciplines. By applying IS experience to the scholar's research questions, we may be able to eliminate much of the drudgery associated with data retrieval and analysis and free the scholar for other tasks. Pilot projects with The Ohio State University (search for rare theater monographs), Pennsylvania State (compilation of dBase bibliographies of Baroque scores and related sound recordings), and an internal project (machine-readable cross-reference of LC classes and subject headings) have had successful results; other pilot projects with Kent State University (serials use) and Colorado School of Mines (ILL use) are pending.

---

Project staff: John A. Bunge, Associate Research Scientist; Nancy Gould, Research Assistant

# Library of Congress Class M Subset: Music

**Project Manager:** Jeanette M. Drone, Postdoctoral Fellow

**Abstract**

Four related studies were completed. First, a two-part study of uniform titles in scores records included an analysis of their currentness and an experimental study to determine the potential for automatically generating AACR2 uniform titles using OCLC Authority File data. Second, a prototype hypermedia music reference system, HyperBach, was developed. Third, a machine-readable file of Library of Congress subject headings for music with corresponding classification data was developed to generate alphabetical and classified listings and a hypertext database. Fourth, a machine-readable file of AACR2 uniform titles, variant titles, and obsolete uniform titles for works of eight prolific composers was developed using data extracted from the Library of Congress (LC) Name-Authority File.

Analysis of uniform titles (MARC tag 240) in selected scores records derived from the OCLC Online Union Catalog for two major composers indicates that work numbers (numerical indicators) are the most consistently and accurately applied element in form/genre uniform titles. To determine the potential for automatically enhancing obsolete or incomplete uniform titles, an experimental study was designed in which AACR2 uniform titles were generated automatically by comparing work numbers in the scores records with corresponding data from the OCLC Authority File.

Scores records for the works of Ludwig van Beethoven were selected as the basis of the study because they represent a wide variety of scoring media and musical forms as well as a complicated numbering system. The design of an algorithm for procedures of this complexity provided valuable data concerning the feasibility of modifying and applying the procedures to the records of other composers. A total of 4,336 records with *Beethoven, Ludwig van* in the main entry heading (MARC tag 100) was extracted from the Online Union Catalog; 4,140 records containing uniform titles were used in the study. Records in the OCLC Authority File for the works of Beethoven were identified, and those with work number data were used to create a uniform title authority file. The uniform title in each bibliographic record was scanned for a work number and, if identified, the number was compared with the work numbers in the uniform title authority file data. When a match was found, the AACR2 uniform title from the uniform title authority file was inserted in the bibliographic record as a supplementary uniform title.

Results of this study indicate that when work number data are available in OCLC Authority File records, bibliographic records with work numbers in the uniform titles can be matched with the AACR2 form to verify the currentness of the data in the bibliographic record uniform title (MARC tag 240). An additional or supplementary uniform title meeting AACR2 standards can be generated automatically and inserted in the bibliographic record.

### HyperBach

To explore the design and engineering issues of integrating text, graphics of musical themes, and audio playback and retrieval, a prototype hypermedia music reference system was developed (Drone 1988). Apple® Computer's HyperCard™ provided database retrieval. The following data were entered for each of 14 works by Johann Sebastian Bach: title, medium of performance (scoring), work number (numerical indicator), references to the collected editions, date of composition, place of first performance, uniform title, and variant titles when applicable. The uniform and variant titles were extracted from OCLC Authority File records. The remainder of the data were verified in Schmieder's (1981) *Bach-Werke-Verzeichnis (BWV)* and *The New Grove Dictionary of Music and Musicians* (1980, s.v. "Bach, Johann Sebastian"). The graphic representations of the thematic materials were entered using Great Wave Software's ConcertWare™ and a Yamaha DX-7 synthesizer.

The HyperBach system demonstrates that it is technically possible to develop an electronic thematic catalog that links cataloging authority data and related reference data. The availability of digitally encoded themes provides new possibilities for developing audio retrieval systems using personal computers or instruments with musical instrument digital interface (MIDI).

### Library of Congress Subject Headings for Music

Subject authority records containing topical subject headings (MARC tag 150) for music and materials about music were extracted from the Library of Congress subject headings in machine readable form (LCSH-mr) and developed as a separate file. The researcher identified and applied suggested class numbers for records lacking LC classification data (MARC tag 053). Using these data, alphabetical and classified listings were generated. Each entry contained a subject heading and classification data. The LC control number (LCCN, MARC tag 010) was included to link the data to the complete records in the online subject authority file. Examples from these listings appear in tables 1.2 and 1.3. Data provided by the researcher are indicated by an asterisk (*).

In addition to being prepared as a two-part printed document, the file of music subject headings was also developed in a hypertext environment using HyperCard. The database contains approximately 10,000 authorized subject headings and 12,000 *see from* tracings (MARC tag 450).

### LC Uniform Titles for Music

The name-authority records for the following composers were extracted from the LC Name-Authority File: Bach, Beethoven, Brahms, Haydn, Mozart, Schubert, Tchaikovsky, and Telemann. For each of these composers the following listings were generated: (1) an alphabetical listing

**Table 1.2** Alphabetical Listing

| LC Control Number | Added Class Number | LC Class Number | Added Subject Heading | LC Subject Heading |
|---|---|---|---|---|
| 85000696 | | MT955 | | Acting in opera |
| 85001063 | * | M2110 | | Advent music |
| 85001246 | | ML1005 ǂb ML1006 | | Aeolian harp |
| 85001246 | | ML1055.A3 | | Aeolian-vocalion |
| . | | | | |
| . | | | | |
| . | | | | |
| 85011470 | | M1200 ǂb M1268 | | Band music |
| ——— | * | ML128.B23 | * | Band music ǂx Bibliography |

**Table 1.3** Classified Listing

| LC Control Number | Added Class Number | LC Class Number | Added Subject Heading | LC Subject Heading |
|---|---|---|---|---|
| 85064504 | | ML430 | | Imitation (in music) |
| 85088807 | * | ML430.5 | | Music ǂx Performance ǂx History |
| 85089012 | | ML431 ǂc History | | Musical notation |
| 85089012 | | ML431 ǂc History | | Tablature (Musical notation) |
| 85089012 | | ML432 ǂc Reform | | Musical notation |

of AACR2 uniform titles given in MARC tag 100, subfield ǂt; (2) an alphabetical listing of variant titles and obsolete uniform titles as recorded in MARC tag 400, subfield ǂt, with cross-references to the AACR2 form in the first listing; and (3) a numerical listing compiled from work numbers identified in MARC tags 100 and 400 or as verified and applied by the researcher. The variant title and work number listings are designed to be used as an index to the listing of AACR2 uniform titles. LC control numbers in each listing provide a link to complete records in the OCLC Authority File. Examples from the three listings for the works of J. S. Bach appear in tables 1.4-6.

**Table 1.4** J. S. Bach: AACR2 Uniform Titles

| LC Control Number | Uniform Title |
|---|---|
| 82096216 | Ach Gott, wie manches Herzeleid (Cantata), ǂn BWV 3 |
| 85367802 | Ach Gott, wie manches Herzeleid (Cantata), ǂn BWV 58 |
| 88045321 | Ach Gott, wie manches Herzeleid (Cantata), ǂn BWV 58. ǂp Ich bin vergnügt in meinem Leiden |
| 81046715 | Ach, ich sehe, itzt, da ich zur Hochzeit gehe |

**Table 1.5** J. S. Bach: Variant Titles/Obsolete Uniform Titles with Cross-references to AACR2 Uniform Titles

| LC Control Number | Variant Titles/Obsolete Uniform Titles |
|---|---|
| 86130918 | Ach, bleibe doch, mein liebstes Leben **SEE:** Lobet Gott in seinen Reichen. ǂp Ach, bleibe doch, mein liebstes Leben |
| 82096216 | Ach Gott, wie manches Herzeleid, ǂn S. 3 **SEE:** Ach Gott, wie manches Herzeleid (Cantata), ǂn BWV 3 |
| 81128610 | Ach, was soll ich Sünder machen? **SEE:** Partite diverse sopra Ach, was soll ich Sünder machen? |

**Table 1.6** J. S. Bach: Work Number Listing

| Work Number | LC Control Number | Uniform Title |
|---|---|---|
| BWV 1 | 82013165 | Wie schon leuchtet der Morgenstern (Cantata) |
| BWV 3 | 82096216 | Ach Gott, wie manches Herzeleid (Cantata), ǂn BWV 3 |
| BWV 4 | 79145908 | Christ lag in Todesbanden (Cantata) |
| . | | |
| . | | |
| . | | |
| BWV 140 | 82111939 | Wachet auf, ruft uns die Stimme (Cantata) |

## References


Drone, Jeanette M. 1988. HyperBach: a prototype hypermedia music reference system. *OCLC Newsletter* no. 173 (May–June): 11–12.

Schmieder, Wolfgang. 1981. *Thematisch-systematisches Verzeichnis der musikalischen Werke von Johann Sebastian Bach: Bach Werke Verzeichnis (BWV)*. Leipzig: Breitkopf & Härtel.

*The New Grove Dictionary of Music and Musicians*. 1980. s.v. "Bach, Johann Sebastian."



Project staff: (Uniform Titles) Jack Cox, Research Assistant; (HyperBach) Mark Crook, Consulting Systems Analyst; Craig Henderson, Research Assistant; Duane A. Rice, Research Assistant; Kurosh Valanejad, Technical Assistant; (LC Subject Headings) Mark Crook, Consulting Systems Analyst; Diane Vizine-Goetz, Research Scientist; Kurosh Valanejad, Technical Assistant; (LC Uniform Titles) Mark Crook, Consulting Systems Analyst


# Managing Large Retrievals


**Project Manager:** Chandra G. Prabha, Research Scientist



**Abstract**

This project examines two approaches to reducing large retrieval sets: (1) the use of language, format, and date qualifiers; and (2) the use of library classification schemes. Transaction logs of users' queries will be run against a database of bibliographic records to determine reduction factors.


Large retrievals pose a variety of problems. Queries that produce large retrieval sets are paradoxical because they take significant machine resources to process but the resultant retrieval sets, because of their large size, discourage users from reviewing all records retrieved.

The problem of large retrievals can be addressed in several ways. In systems that facilitate keyword searching and provide Boolean operations, users can add terms to restrict searches. Alternatively, a system could order search results according to user specifications or rank retrieved records by the frequency of search terms appearing in each record. Another approach is to qualify search queries by date, language, or format of publication. Many systems providing such qualifications do not calculate or inform the user of the expected reduction in the size of the retrieval set based on the qualifications. The Managing Large Retrievals (MLR) project will concentrate initially on two approaches to reducing large retrieval sets: (1) the use of language, format, and date qualifiers; and (2) the use of library classification schemes.

This project explores to what extent qualifiers reduce retrievals for a given class of predefined queries. For in-

stance, for queries containing a single term can it be predicted that the set will reduce to 53% when the date-range qualifier *1970 or later* is invoked?; to 63% when the language qualifier *English* is invoked?; to just 6% when a combination of qualifiers (e.g., date range, language, format, and the presence of the Library of Congress classification) is invoked?

Another objective is to examine the use of Library of Congress and Dewey Decimal classification schemes for reducing large retrieval sets. Theoretically, library classification schemes collocate books on similar topics. A retrieval set of 1,000 matches to the search query *plants* may represent items in different LC classes (table 1.7).

**Table 1.7** Scatter of Query *Plants* by LC Class

| LC Class | Number |
|---|---|
| Science (Q) | 450 |
| Geography/Anthropology (G) | 100 |
| Agriculture (S) | 200 |
| Technology (T) | 50 |
| Other | 200 |
| Total | 1,000 |

By presenting a breakdown of matched records by Library of Congress (or Dewey Decimal) classes, users are alerted to potentially unwanted search results. Users could select classes that match their interests as an early step in the process of refining the query.

### Method

The OCLC New Systems Reference Service database, currently in development, will be used in the MLR project. This database is expected to contain about 18 million records and permit keyword searching of subject, title, and other fields. The size and heterogeneous character of this database make it especially appropriate for exploring the use of LCC or DDC schemes to help refine search queries. Search queries for this investigation will be drawn from transaction logs of current systems. Queries obtained this way reflect users' actual information needs and searching patterns. The University of Illinois has provided a transaction log of searches performed on its Full Bibliographic Retrieval System during February 1988. We will also use transaction logs from other library systems.

For the University of Illinois data, a computer program indicates the number of matches for each query submitted to the system. Because the log contains hit data for all queries, threshold levels can be increased incrementally to define a large retrieval set. Plans are under way to sort the extracted search queries by types of searches: subject, author, title, and author/title. These queries will be further grouped by the following criteria: number of terms in a query; number of *AND, OR,* and *NOT* operators; and combinations of terms.

A computer program will batch process these sorted queries in the New Systems Reference Service database. We will study the reduction obtained by applying qualifiers, and estimate a reduction factor for various classes of queries.

### Pilot Phase

The MLR project will be conducted in two phases. In the pilot phase a small sample of queries will be used to investigate all aspects of the study. Based on findings of this pilot phase, necessary changes and refinements will be made for the subsequent large-scale study. Using a random sample of 16,843 bibliographic records, .001% of the OCLC Online Union Catalog at the time the sample was extracted, the extent to which selected values of qualifiers (date range, language, format, and classification schemes) reduce the sample set was studied. Results of this investigation will serve as a benchmark for comparing reduction results when qualifiers are applied to actual test queries.

A user could select any attribute of a bibliographic record to restrict search results. Some attributes, however, are encoded more frequently than others. Format information is present for nearly 100% of the sample set; date and language of publication are present in 98%. However, cataloging source is present only in 55% of the records. The greater the frequency of a particular attribute, the more useful it is as a qualifier. For this reason, language, date, and format qualifiers were selected.

Qualifiers were applied singly and in combinations. Reduction results when qualifier values were applied singly appear in figure 1.12. For instance, the format qualifier *books* reduced the retrieval set to 79% (column A). Given the largely monographic character of the database, one would not expect the set to reduce further for this qualifier value. Therefore, the format qualifier *books* should be used with other qualifiers.

The set reduced to 25% (column D) when the date qualifier *1980 or later* was applied. If recency of publication is a criterion for selecting citations, then the application of recent date-range qualifiers could be useful for reducing large retrievals.

Reduction results obtained by applying combinations of qualifiers appear in figure 1.13. Date-range qualifiers with values *1970 or later* and *1980 or later* were first applied. The set reduced to 53% and 25%, respectively (columns A and B). Subsequently, the language qualifier *English*, the format qualifier *books*, and classification qualifier *Library of Congress classification number* were applied successively. This combination reduced the retrieval set to 19% and 8% respectively (columns G and H). A potential reduction of over 90% was achieved by qualifying searches to English-language monographs published after 1980 with a Library of Congress classification number. If most users search for items meeting these specifications, then perhaps new ways of processing could be explored to save computer resources.

The pilot phase will be completed after a small query set is searched on the New Systems Reference Service database. Based on the findings of the pilot phase, the research questions and methods will be refined for the large-scale study.

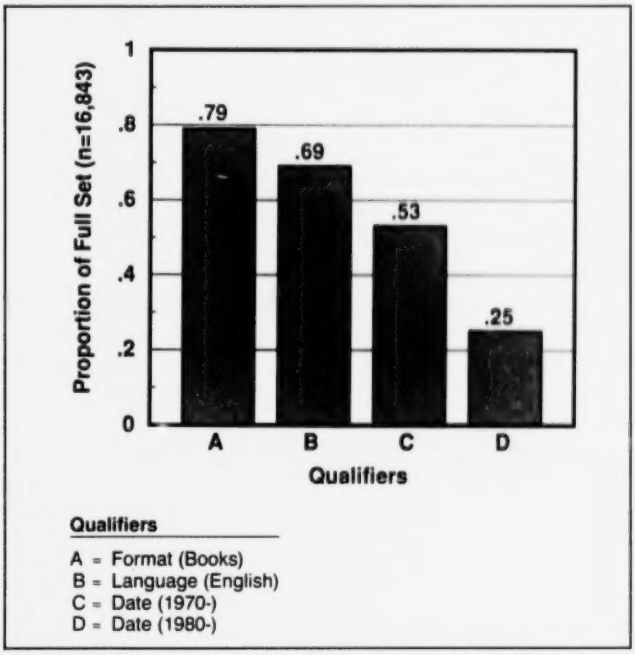



**Figure 1.12** Retrieval Set Reduction with Single Qualifiers

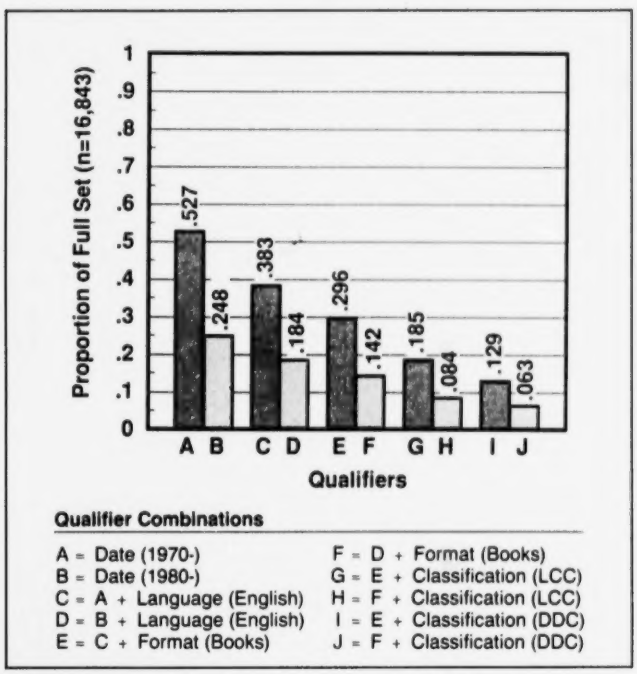



**Figure 1.13** Retrieval Set Reduction with Combined Qualifiers

# Nonfiction Book Use

**Project Manager:** Chandra G. Prabha, Research Scientist

**Abstract**
Nonfiction book use studies conducted in public and academic libraries are described and compared. Findings support the assumptions that users generally read nonfiction books to obtain information, and that they read nonfiction books in parts. Data on methods of selecting nonfiction books and finding information are presented. Results of these book use studies support development of information systems that provide access to the text of nonfiction books.

Library books provided in an electronic information system would never be checked out, off the shelf, in a reserve collection, at the bindery, or otherwise unavailable for use. Users could be certain of using any book on the system. Telecommunications systems and computer networks could even provide users access to the information system from homes or offices. Users could find information immediately when and where they need it.

Such a scenario, however, assumes certain patterns of information seeking behavior: (1) library users read nonfiction books generally to obtain information, and (2) they generally read nonfiction books in parts. Two book use studies examined these assumptions, first in a public library and later in an academic library. Data on how users select nonfiction books, and how they find information within a book, were obtained and analyzed. In this report the findings from the public and academic library studies are compared and some conclusions are drawn.

Nonfiction books include dictionaries, telephone directories, handbooks, maps, and much more. In libraries such materials are regarded as reference books and are generally not loaned for use outside the library. However, most libraries permit users to borrow nonfiction books that treat subjects in some degree of depth. This category of nonfiction books was the focus of our study.

## Method

The Public Library of Columbus and Franklin County and The Ohio State University Library system participated in our book use studies. In the former we collected data at four branch libraries representing different types of neighborhoods and the main library; in the latter we collected data just at the main library. We used proportionate sampling in the public library and cluster sampling in the academic library. A full discussion of how libraries were selected and criteria used to determine the number of interviews is presented elsewhere (Prabha, Rice, and Cameron 1988).

During a pretest phase of the public library study, an interview script was developed. A modified version of the script used in the public library was pretested in the academic library and refined. We interviewed users in the public library during February and March 1986 and in the

academic library in February and March of 1987. To improve recall, interview questions centered on the books users remembered best.

### Participant Profile

An analysis of 347 participants at the public library showed that a nonfiction borrower at a public library is likely to be a male (58.4%), or between 25 and 44 years old (59.1%), or at least partially college educated (90.9%), or employed full-time (56.5%), or in the household income range $10,000–$20,000 (43.7%).

In the academic library, 426 nonfiction book borrowers were interviewed. Most were undergraduates (44.7%); graduate students comprised the next largest group (38.6%). Faculty represented 10.4%; continuing education students, staff, and others constituted the remaining 6.5% (percentages are rounded).

### Findings

The data largely support the assumption that nonfiction books are read generally to obtain information. Interviewees were asked to think of information as one or more ideas or facts. In the public library, 94.6% of the users were looking for information; only 5.4% stated explicitly that they were not looking for information. The corresponding percentages for the academic library were 97.6% and 2.4%. We sought to determine if users read nonfiction books to obtain specific information (table 1.8). We did not define what constituted "specific" information; answers reflect users' judgments. In both library settings, at least 50% of users described the information they were looking for as specific. It appears that academic library users are more likely to be looking for specific information than are public library users.

**Table 1.8** Nonfiction Book Use for Specific Information

| | Library Type | |
|---|---|---|
| Type of Information | Public n = 154* (%) | Academic n = 426 (%) |
| Specific | 50 | 75.2 |
| General | 50 | 24.2 |

$X^2 = 31.54$, $p = .005$, $df = 1$
* Data from main library only

The second assumption, that nonfiction books are read generally in parts, was supported but not so strongly as the first assumption. At the academic library 46.2% of users, in contrast to 31.7% at the public library, read about one-fourth of the book or less (table 1.9). For users reading less than one-fourth of the book the figures are 24.9% and 16.2%, respectively (not shown in table). This trend reverses for almost all or all of the book.

**Table 1.9** Amount of Book Read

| | Library Type | |
|---|---|---|
| Amount of Book Read | Public n = 154* (%) | Academic n = 426 (%) |
| About ¼ | 31.7 | 46.2 |
| About ½ to ¾ | 22.7 | 27.4 |
| Almost all or all | 45.5 | 26.4 |

$X^2 = 19.76$, $p = .005$, $df = 2$
* Data from main library only

Can a user who reads one-fourth or less of a book be viewed as a target user of nonfiction books in an electronic mode? How do users who read short sections of a book find information that is relevant to their needs? In the academic library 29.4% of users found the information in one part of the book. While 30.2% found information in all parts of the book, the largest group of users, 40.5%, said that the information they sought was in different parts of the book. The corresponding percentages of the public library study are 6.7%, 33.6%, and 59.7%, respectively. Clearly these data indicate further research is needed to learn how users navigate through books.

**Selecting nonfiction books.** Of academic library users, 46.2% knew about the book before coming to the library (specific-item users) compared with 28.6% in the public library; 37.5% selected the book at the catalog (catalog users), the corresponding percentage for public library is 21.4%; 15.7% selected the book at the book shelves (shelf browsers) whereas 50% selected likewise in the public library.

**Finding information in nonfiction books.** Having obtained a book, how do users locate information? In the academic library environment, 61% used the table of contents or book index compared with 41.6% in the public library. The use of tables of contents or book indexes may assume a greater importance in electronic documents. We sought this data so designers of full-text systems of nonfiction books can develop appropriate finding and navigational aids.

**Users.** Analysis of responses to four key questions posed to academic library users (book selection method, the context in which the book was used, amount read, and the features used for finding specific information) showed that 23.2% needed the book for school, read one-fourth of the book or less, and used the table of contents or book index.

A similar analysis of the data of the public library grouped users according to a series of decisions they made in their information search. This analysis showed that eight principal paths were taken by 70% of the users and

provided an empirical method for grouping nonfiction book users: true browsers, 26.3%; information-seeking browsers, 17.7%; known-item users, 13%; and information-seeking catalog users, 11.8% (percentages are rounded).

### Conclusions

Three out of four users in the academic library were looking for specific information. In the public library two out of three were looking for specific information. In the academic library nearly half of all the users read one-fourth of the book or less; in the public library one-fourth of the users read one-fourth of the book or less. Given these findings, academic library users may be a better audience, at least initially, than public library users for electronic access to full-text of nonfiction books.

### References

Prabha, Chandra G., Duane A. Rice, and David Cameron. 1988. Nonfiction book use by academic library users. Research report no. OCLC/OR/RR-88/1. Dublin, Ohio: OCLC.

Project staff: John Bunge, Associate Research Scientist; David Cameron, Research Assistant; Duane Rice, Research Assistant

## Scholars Cross-Reference System

**Project Manager:** Paul B. Kantor, Visiting Scholar

**Abstract**
The Scholars Cross-Reference (SXR) system is a novel approach to collecting, enhancing, and using bibliographic relationships not manifest in text per se. The major system features and guidelines for prototyping have been developed. In addition to practical value in bibliography development and collection management, the SXR system is a value-added service to online catalog users and makes possible new research into the structure of knowledge.

The Scholars Cross-Reference (SXR) system enables scholars and researchers to collate and use bibliographic information otherwise unavailable. The user of an online catalog can browse through a file of related materials and download information. The browsing file is built by scholarly users of the catalog who suggest that certain items are related. Because this file is not built by conventional techniques such as automatic text processing or expert cataloging, it has the potential to reveal unsuspected links among items that are apparent only to those knowledgeable in the subject area.

In addition to browsing, users can upload materials such as personally developed bibliographic lists. Such lists are added to those maintained by the Scholars Cross-Reference system. In this way the SXR files can grow and change. Techniques for managing these files must include, at a minimum, the automatic identification of citations that are likely to refer to the same work, maintenance of frequency counts (to identify the items most heavily cited as being related to a given item), and a weeding mechanism.

When in full operation, the Scholars Cross-Reference system will maintain lists of related items that are associated to catalog entries for which users have suggested related materials. In practice, this will be a small fraction of the total catalog holdings, most of which are unused for years at a time.

### System Structure

There are two major elements to the system structure, a gateway system and a locally resident user interface. The main database of bibliographic information may be local or remote. If the users are a group of scholars at a university and the main database is the OCLC Online Union Catalog, for example, the gateway system could be located either at the campus computer center or at OCLC. If the main database is the online public access catalog at a university, it makes sense for the gateway system to be maintained at the library.

The gateway maintains files of related materials labeled by the identifier (e.g., accession number or OCLC control number) of the item to which the contents of the file are related. The gateway system functions in either transparent or interactive mode. In transparent mode it relays traffic from the users to the main database. It can be "wakened" from the transparent mode either by a signal sent by the user or in response to a chosen trigger in the traffic coming from the main database. One example of a trigger is the transmission of a full bibliographic record.

In interactive mode, the gateway system conducts a conversation with the user, provides access to and downloading of the files of related items, and accepts additional uploaded information for those files. The gateway system should also have a module to monitor traffic and to provide information for the management of the system.

The second component of the system, the user interface, resides at each user's terminal. The terminal could, for example, be an personal computer using a suitable communications package under the control of the SXR system user interface. The user interface also functions in two modes, transparent and interactive. In the transparent mode it relays conversation from the keyboard to the communications board and from the communications board to the screen. In the interactive mode it manages browsing, selecting, downloading, and uploading files. The user interface system should also contain a module to monitor activities that would support research and development of the Scholars Cross-Reference system itself.

### Development

During the second half of 1987 an effort was initiated to prototype the user interface and the gateway system using available software. Most of the functionality of the user system has been simulated using the DESQview™ windowing software, the Pro-Cite™ bibliographic file manager, the SmartKey key stroke multiplier, and the EDIX™ word processor. The prototype can exhibit each of the functionalities in a rudimentary form.

Prototype development of the gateway system was initiated using the Crosstalk® communications package and the DESQview windowing system. The current version of the Crosstalk system manages two communication sessions simultaneously. Thus, the gateway system can be connected via telephone lines to both the user and the main database.

### Practical Potential

In addition to providing a value-added service to the users of an online catalog, the Scholars Cross-Reference system has potential practical value for library system managers. For purposes of collection management, the linked files could be scanned periodically to identify frequently recommended items that are not in the library's collection. Additionally, library staff could compile bibliographies that would draw not only on conventional bibliographic resources or personal knowledge and expertise but also on the suggestions provided by scholars and researchers. Such bibliographies would be timely and relevant to those the library serves.

### Research Potential

The Scholars Cross-Reference system facilitates three kinds of research. The system serves as a basis for research on user perceptions, patterns of use, and interface quality. This is necessary to develop the most useful tool for libraries and scholars. A second research topic addresses the impact of the SXR system on scholarship itself. This requires longitudinal studies to collect indicators of scholarly effectiveness such as productivity and student evaluations. Finally, the SXR system makes possible a new kind of research into the structure of knowledge, which is presently studied by patterns of co-citation and similar manifest characteristics of text.

# Selecting Bibliographic Records

**Project Manager:** John A. Bunge, Associate Research Scientist

**Abstract**

This project developed and tested methods to determine optimal sets of bibliographic records to be extracted from the OCLC Online Union Catalog for storage on compact discs for use in cataloging and retrospective conversion. Of selection criteria examined, recent use was the best predictor of future use. A subset of 500,000 records for books dated 1982 and later is projected to satisfy 89.2% of current cataloging needs.

This project continues work to extract optimal subsets of bibliographic records from the OCLC Online Union Catalog. Records will be stored on compact discs for cataloging and retrospective conversion activities using microcomputers.

### Method

All cataloging transactions by all libraries using the OCLC Online System were recorded for a two week period in March 1988. During this period 1.2 million transactions occurred on 719,990 records or 4.1% of the 17.5 million records in the database at that time. Of these records, approximately 532,000 were used for current cataloging and 319,000 were used for retrospective conversion. The task of the project was to learn how to differentiate records receiving high use from the remainder of records in the database.

On the first day of sampling 192,330 records were randomly drawn from the institutional holdings bit (IHB) file to preserve first-day holdings information (this file retains holdings data associated with bibliographic records). The records were then further processed to produce data files containing number of hits and selected other bibliographic information. Absolute time was then transformed into relative time variables. The date of last use, for example, taken from the IHB file, was changed to days elapsed since last use. Similarly, Date1 was changed to age in years (e.g., Date1 = 1988, age = 0; Date1 = 1987, age = 1). This was done to make the analyses more generally applicable to records as time goes on.

### Decision Trees

The objective was to use the IHB count, elapsed time, age, record type, and bib level to project the future use of each record. Records were classified into two groups each for current cataloging and retrospective conversion activity using the Classification and Regression Trees (CART) method (Breiman et al. 1984): records with no activity (no hits) and records with some activity (one or more hits).

The CART method attempts to find "tree-structured classifiers" on the basis of IHB and bibliographic data. Six cases were examined: (1) current cataloging for all records,

(2) current cataloging for books with imprint 1982 or later, (3) current cataloging for all records except books with imprint 1982 or later, (4) retrospective conversion cataloging for all records, (5) retrospective conversion for books with imprint 1982 or later, and (6) restrospective conversion for all records except books with imprint 1982 or later. In each case CART constructs a decision tree for optimal record selection. A record enters the tree at the top and, based on a sequence of criteria, a decision is made to select the record or not (figure 1.14). Records end up in one of the terminal nodes. The records are selected if the node has a number and rejected if the node has an X.

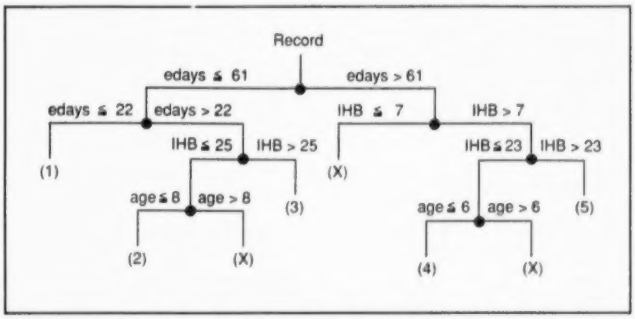



**Figure 1.14** CART Decision Tree. Current cataloging for books with imprint 1982 or later (edays = elapsed days). Optimal subsets are numbered 1–5.

In this case, all five optimal subsets contain 3,029,509 records and account for 90.2% of all current cataloging. These results suggest that recent use is a good indicator of future use. Analysis of record use reveals that for books with imprint 1982 or later 96% was current cataloging; 4%, retrospective conversion. A set of 500,000 records selected on the basis of date of last use would satisfy 89.2% of current cataloging for books with imprint 1982 or earlier.

## References

Breiman, Leo, et al. 1984. *Classification and Regression Trees*. Belmont, Calif.: Wadsworth International Group.

# 2

# External and Collaborative Research

OCLC research activities extend beyond the walls of the Center in Dublin, Ohio, by supporting research projects outside of OCLC and collaborating on joint projects.

As a founding member of the Applied Information Technologies Research Center (AITRC), a nonprofit consortium of scientific, educational, and business organizations dedicated to the development of computer information systems, OCLC gains access to recent technologies, products, and services.

AITRC members collaborate on applied research projects that correspond to OCLC interests including search technologies, intelligent user interfaces, information representation, and user behavior.

OCLC is also a sponsor of the Massachusetts Institute of Technology Media Laboratory, a unique interdisciplinary laboratory for advanced study in new information technologies. The Media Laboratory supports innovative applications of electronic and digital communication technologies to enhance the creation, storage, and delivery of multimedia information. Research results are nonproprietary and available to sponsoring agencies.

OCLC also supports specific research projects by contributing funds, staff, or computing resources. A description of research to increase the accessibility of Library of Congress subject headings in online bibliographic systems appears in this section (page 32). This project is funded jointly by the Council on Library Resources, The University of Michigan, and OCLC. Complete interim findings describing the characteristics of topical subject ►

headings and geographical names in the machine-readable Library of Congress subject headings (LCSH-mr) are available from OCLC.

Project Mercury, a joint effort by Carnegie Mellon University and OCLC, represents a major commitment by both parties to develop and implement a distributed electronic library (see page 36). The project will unify communication, storage, access, and information display technologies to deliver information to the user's desktop workstation.

Other projects reported in this section range from automatic indexing and term weighting to development of an intelligent interface and a detailed study of OCLC bibliographic records for Latvian materials.

Extramural activities add a vital dimension to OCLC research and complement Office of Research activities, and whenever possible methods and findings of these projects are incorporated in OCLC research projects. The cooperative interaction among researchers and institutions is mutually beneficial and encourages research not otherwise possible.

# Automatic Phrase Construction for the Representation of Text Content

**Principal Investigator:** Gerard Salton, Professor, Department of Computer Science, Cornell University, Ithaca, New York

**Abstract**

This project is designed to generate complex identifiers for the content representation of natural language texts. Such identifiers can serve as entries in book indexes and may lead to improved retrieval effectiveness in automatic text retrieval systems. The phrase construction system includes both syntactic and semantic features.

Complex content identifiers are descriptions consisting of combinations of nouns and adjectives such as *information retrieval systems*. A standard process for generating complex identifiers is to perform a syntactic text analysis of a document, or document excerpt, followed by the identification of certain nominal constructions that are present in the text. Thus, for a text sentence such as

> with the exception of the development of a negative dictionary all system operations are completely automatic

the phrase identifiers might be *development exception* from *exception of the development*, *dictionary development* from *development of a dictionary*, and *system operations* from *system operations*.

A phrase identification process based only on syntactic considerations generates many useful phrases, but also many extraneous ones. The sample sentence may reasonably be identified by the phrase *dictionary development* but not by *development exception*. In practice, the presence of questionable content identifiers causes false matches between query statements and stored documents and reduces retrieval effectiveness. Available experimental evidence indicates that for certain document collections the use of complex phrases does not produce improvements in retrieval effectiveness over much simpler single-term content identifiers.[1]

An evaluation of the currently available phrase generation system indicates that semantic features must be used in addition to the syntactic processes to improve the phrase generation method. Furthermore, positional criteria that take into account the text location of the phrase components may also prove useful. The following refinements in the phrase generation methods are currently under investigation:

1. Adjectives may be deleted before accepting complex phrases for content identification. Thus *large data block* is simplified to *data block*, and *conventional permutation* to *permutation*. Technical adjectives should, however, be kept intact; thus, *cryptographic transformation* must not reduce to *transformation*.
2. Common nouns, included on a special list, may be deleted before complex phrases are accepted as content identifiers. This includes words such as *system*, *method*, *technique*, and so on. The phrase *data enciphering system* is thus reduced to *data enciphering*.
3. Phrase location may be taken into account. Phrases derived from document titles, or headings and captions, may be given special treatment.
4. Phrase appearance must be taken into account in generating phrase identifiers. Phrases that include capitalized or italicized components are kept intact without term deletion. The phrase *Data Encryption System* is not reduced by deletion of *System*.

Some examples illustrating the phrase transformation rules appear in table 2.1.

The refined phrase indexing package is being implemented now, and experiments are planned to test the new system under operational conditions. Additional work also

under way relates to the identification of duplicated, or partly overlapping, phrases, the automatic identification and placement of phrase cross-references of the kind used in ordinary book indexes, and the treatment of prepositions and conjunctions that appear in nominal constructions.

**Table 2.1** Phrase Transformation

| Phrase Extracted from Document Text | Corresponding Indexing Phrase |
|---|---|
| Data Encryption System | Data Encryption System |
| data encryption system | data encryption |
| transformation of blocks of data | data block transformation |
| conventional permutations and substitutions | conventional permutations conventional substitutions |
| large data blocks | data blocks |
| key information | key |

### Notes

1. See, for example, J. L. Fagan. 1987. *Experiments in phrase indexing*. Technical Report 87-868 (Ithaca, N.Y.: Cornell University Department of Computer Science).

# Extensions to the Advanced Interface Management Project

**Principal Investigator:** Niall Teskey, Deputy Head, Information Technology Research Institute, Brighton Polytechnic, Brighton, U.K.

**Abstract**
A demonstration system shows that an intelligent assistant can extend the functionality of a full-text information retrieval system. The system extends the user interface, developed using the D language, to provide assistance in two common searching scenarios.

The Advanced Interface Management project developed a series of interfaces to various types of information retrieval systems. The interfaces are implemented using the D language which was developed at OCLC. The D language uses the window, icon, and mouse programming style to develop highly interactive interfaces. This project investigates ways in which the interface can adapt itself and respond to the user's performance or search results. The Poplog system is used to provide an artificial intelligence (AI) environment that runs on the same hardware (Sun Workstations) as the D language, can be linked to other systems, and allows for the rapid development of prototypes.

The project has three stages: (1) identify areas of the interface that would benefit from the proposed AI extension, (2) develop prototypes to evaluate methods of data transfer between the interface and the Poplog system, and (3) build a prototype system to demonstrate proposed extensions.

### Proposed Extensions

While the D interpreter runs the interface, a Poplog process runs in the background monitoring the conduct of the search and providing assistance when necessary. This background process is called an intelligent assistant. The assistant can operate in two modes, request and monitor. In request mode, if the D program needs help, it asks for help from the assistant. In monitor mode the assistant is constantly watching the progress of the D program and, if it detects irregularities, it interrupts the program and attempts to correct the problem. Interfaces to information retrieval systems need both modes; examples of each are given below.

**Request mode.** If a search retrieves too many or too few hits, the interface detects the problem and requests help from the assistant. When a search fails, the D program passes the search terms and hit list to the assistant for further processing. The assistant tries various modifications to the search and returns either a modified search and hit list or a message that the new searches had failed. The interface then resumes its normal processing or switches control to some error display state.

**Monitor mode.** If a user repeatedly issues the same command and gets an unsatisfactory system response, this may indicate that the user does not fully understand correct command usage. By monitoring the actions of the interface, the assistant identifies sequences of commands and attempts to discover why the user is having difficulty. The interface provides the assistant with details of the commands that the user entered. The assistant analyzes the commands, interrupts the interface, and switches into a tutorial mode to explain the problem to the user.

### Feasibility Studies

Two types of data transfer between the interface system and Poplog are needed: (1) synchronous data transfer initiated by the interface in request mode, and (2) asynchronous data transfer initiated by Poplog in monitor mode.

The simplest method of synchronous transfer of data between D and Poplog is via UNIX files. Both languages have system and file access commands, although those in D are less comprehensive than those in Poplog. The system uses two files, D__in and D__out, for input and output from D. When the system fails to find search terms, it writes a message to the file D__out rather than warning the user. The program then waits for a message to come back in the file D__in and displays this message. The Poplog sys-

tem is running in parallel waiting for a message to be written to the file D_out. As soon as this is detected, the Poplog system generates a reply and writes it to the file D_in for the D program to access.

Developing methods of asynchronous data transfer between D and Poplog is more difficult. The D interpreter provides a complete log of each user session. Because this log file is created dynamically during the search session, it can be examined to monitor the progress of the search. The Poplog system runs in the background and monitors this file, waiting for new records to be added. If the Poplog system detects an abnormality that can be corrected, it provides corrective information in the file D_cmd. The D interface periodically checks this file and invokes any commands that Poplog has placed there.

### Prototype System

A prototype system has been developed to demonstrate the use of these techniques. The prototype is based on a demonstration system implemented in D using the full text of a book. Extensions added to the system handle the following situations: (1) a single search term retrieves too many hits, and (2) the user repeatedly gets few hits.

If a search returns more than some predetermined maximum number of hits (currently 30), the system passes the hit list to the Poplog assistant for further processing. The assistant first looks for occurrences of the search term in the table of contents and book index. Occurrences of the term, if any, are offered to the user. For example, the term *international* retrieves too many items but it appears in the index as *international competition* and *international marketing*. These terms are offered to help refine the search. If the search term does not appear in the table of contents or index, the Poplog system attempts to locate areas of the book that are relevant to the search term by combining the individual hits into ranges of relevant paragraphs which then are presented to the user. If there is still too much information, a random selection is presented.

With any interface a user may repeatedly enter commands that are not performing in the way expected; system response is less than optimal. For example, in the demonstration system terms are implicitly *AND*ed together; if a user repeatedly searches for multiple terms and retrieves few or no hits, the user may not realize the difference between the Boolean operators *AND* and *OR*. The Poplog assistant constantly monitors the performance of the interface and tracks the number of searches of each particular type. When the frequency of use of any poor search strategy exceeds a certain threshold, the assistant interrupts the user interface to request a help display. At present this consists of some standard help files incorporated into the interface but could include specific help from the Poplog assistant.

### Conclusions

The demonstration system has shown that an intelligent assistant can be used to extend the functionality of an information retrieval interface. Two possible extensions have been studied and several methods of implementing them have been investigated. The next stage is to establish what types of extensions are most suitable for particular applications such as full-text databases or online catalogs. The existing D and Poplog systems with some changes are suitable for this stage. The next stage, evaluating the interface and its extensions, requires that the two systems be integrated into a more efficient and robust system suitable for field tests.

---

Niall Teskey conducted his research at OCLC as Visiting Research Fellow funded by the British Library.

# Increasing the Accessibility of Library of Congress Subject Headings in Online Bibliographic Systems

**Principal Investigators:** Karen Markey, Assistant Professor, School of Information and Library Studies, The University of Michigan, Ann Arbor; and Diane Vizine-Goetz, Research Scientist, OCLC, Dublin, Ohio

**Abstract**
OCLC and The University of Michigan have joined in a collaborative research project to investigate automated techniques to link subject terms entered by online catalog users with the library's controlled vocabulary. This report describes the characteristics of authority records for topical subject headings and geographical names in the machine-readable Library of Congress subject headings (LCSH-mr). The project is supported by the Council on Library Resources, OCLC, and The University of Michigan.

This report is an account of the preliminary statistical analysis of LCSH-mr conducted in Phase I of the project. The statistical analysis answers the first research question:

> What are the characteristics of authority records for topical subject headings (MARC tag 150) and geographical names (MARC tag 151) in the machine-readable Library of Congress Subject Headings (LCSH-mr)?

### Characteristics of Subject Authority Records (MARC Tags 1xx)

LCSH-mr contains 160,706 subject authority records with 32 different types of fields and 152 different types of subfields. The average number of variable-length fields per LCSH-mr record is 6.43 and each field averages 21.07 characters. The maximum number of variable-length fields used

in a record is 432 for the established subject heading *Indians of North America*. The average length of variable-length fields is 135.56 characters. The longest record is for the established geographical name *Japan—Antiquities* which has 14,145 characters.

The majority (80.39%) of LCSH-mr records have been established for topical subject headings (MARC tag 150); 11.96% of records have been established for geographical names (MARC tag 151).

### Broader and Narrower Terms in LCSH-mr Records

Printed editions of LCSH contain *see also* references (designated sa) and *see also from* tracings (designated xx). LCSH-mr, however, contains only fields for *see also from* tracings (MARC tags 5xx). Because one of the concerns of the project is integrating LCSH-mr and its term relationships into online bibliographic systems, *see also* references (local tags 9xx) were generated from *see also from* tracings and added to records. *See also from* tracings (MARC tags 5xx) were designated as broader and related terms and *see also* references (tags 9xx) as related or narrower terms. Except for personal name *see also* tracings, most *see also from* tracings are broader terms. Similarly, except for personal names, most *see also* references are narrower terms. Overall, 75.75% of LCSH-mr records do not bear narrower terms. When narrower terms occur in LCSH-mr, an average of 3.64 narrower terms occurs per record. For topical subject headings (MARC tag 150) and geographical names (MARC tag 151), 2,750 records (1.86%) have more than ten *see also* references. The subject heading *Indians of North America* has 416 *see also* references for topical subjects. Narrower terms are an important component of subject authority records because they guide library catalog users from the general terms they consult in the catalog to more specific terms that better represent their topics of interest.

### Cross-Reference Structure in LCSH-mr Records

*See from* tracings (MARC tags 4xx) lead from unestablished headings to established ones. For example, the topical subject heading *AIDS (Disease)* has the following *see from* tracings:

450 Acquired immune deficiency syndrome
450 Acquired immunodeficiency syndrome
450 Acquired immunological deficiency syndrome

For topical subject heading records, 57.24% have no *see from* tracings; for geographical name records, 72.36% have no see from tracings (MARC tags 4xx). When records for established subject headings or geographical names bear no *see from* tracings, the burden of searching is on catalog users who must know the correct form of heading or name.

**Orphan headings.** The term *orphan heading* is used to describe a subject heading that is not accompanied by one or more *see also* references, *see also from* tracings, or LC class numbers. One definition of an orphan heading is a heading that has no parents or broader terms. This is the most stringent definition of an orphan heading, and with respect to this definition, 39.74% and 56.67% of topical subject headings and geographical names, respectively, are orphan headings. An example of such an orphan topical subject heading with no broader terms is *Cell division* which is accompanied only by the following related (RT) and narrower (NT) terms:

RT Cell proliferation
NT Amitosis
Chromosome replication
Contact inhibition (Biology)
Cytokinesis
Karyokinesis
Meiosis
Mitosis

### Characteristics of Topical Subject Heading Records (MARC Tag 150)

LCSH-mr contains 128,367 records bearing established headings for topical subjects (MARC tag 150). With respect to cross-reference fields, 42.48% of records bears one or more *see from* tracings for topical subjects (MARC tag 450), 61.25% bears one or more *see also from* tracings for topical subjects (MARC tag 550), and 29.74% bears one or more *see also* references for topical subjects (local tag 950). Fewer than 3% of topical subject heading records contain cross-reference fields for geographical names (MARC tags 451, 551, and tag 951). Established headings for topical subjects (MARC tag 150), topical subject *see from* and *see also from* tracings (MARC tags 450 and 550), and topical subject *see also* references (tag 950) contribute the greatest number of characters per topical subject heading record, in part because they are frequently occurring fields in these records.

LC classification number fields (MARC tag 053) occur in 43.02% of records, and scope notes (MARC tag 680) occur in 2.41% of records. Scope notes, general explanatory *see also* references (MARC tag 360), and source data found and source data not found fields (MARC tags 670 and 675) contribute the greatest number of characters to topical subject heading records; however, these fields occur in fewer than 5% of records so their average contribution of characters to records is much smaller than their contribution per occurrence.

### Characteristics of Geographical Name Records (MARC Tag 151)

LCSH-mr contains 19,218 records bearing established headings for geographical names (MARC tag 151). LC classification number fields (MARC tag 053) are contained in 10.29% of geographical name records; .97% contain scope notes (MARC tag 680). These percentages are considerably less than those for topical subject headings. Percentages of *see from* tracings, *see also from* tracings, and *see also* refer-

ences for topical subjects (MARC tags 450, 550, and tag 950, respectively), are lower than those in topical subject heading records generally. However, geographical name records contain larger percentages of *see from* tracings, *see also from* tracings, and *see also* references for geographical names (MARC tags 451, 551, and tag 951, respectively) than topical subject heading records. Fields for geographical names, *see from* tracings, *see also from* tracings, and source data found contribute the most characters per record.

### Unsubdivided and Subdivided Established Headings

Topical subject heading records (MARC tag 150) and geographical name records (MARC tag 151) were distributed into three separate and mutually exclusive record categories: (1) unsubdivided headings, (2) once-subdivided headings, and (3) multisubdivided headings. The majority (70.48%) of topical subject headings are unsubdivided; 25.90% of subject headings are accompanied by one subdivision and 3.62% are accompanied by two or more subdivisions. Examples of unsubdivided, once-subdivided, and multisubdivided topical subject headings are

150 Gardening

150 Gardening ǂx Equipment and supplies

150 Gardening ǂx Equipment and supplies ǂx Marketing

For geographical names, 47.72% are unsubdivided; 38.67% are multisubdivided, and 13.61% are once-subdivided.

### Future Research Plans

Generally, the analysis of LCSH-mr increases our understanding of this file of subject authority records. Conclusions and recommendations about searching and browsing subject authority records are not made in the project's interim report. They will be made in a final report after additional analyses are performed on LCSH-mr in comparisons with assigned subject headings from a bibliographic file and with subject access points from online catalog transaction logs. The preliminary analysis of LCSH-mr is necessary to facilitate comparisons in later phases of the project.

Project staff: Tony Ershadi, Programmer/Analyst; Joseph Reilman, Technical Assistant, OCLC

# Latvian Bibliographic Records in the Online Union Catalog

**Principal Investigator:** Inese Auzina Smith, Lecturer, Department of Library and Information Studies, Loughborough University of Technology, Leicestershire, U.K.

**Abstract**

This project developed a methodology to assess subject resources for a narrow subject speciality in the OCLC Online Union Catalog. The methodology was then applied as a case study to one subset of Baltic bibliography, Latvian records. These records were analyzed according to language, type of record, bibliographic level, cataloging source institution, holdings, subject distribution, imprint, and extensivity. Several aspects of database quality relative to Latvian records were also investigated.

Latvian records in the OCLC Online Union Catalog were analyzed in detail as part of a project to develop techniques to examine small, specialized data subsets.

### Methodology

Latvian records were identified by searching the OCLC Online Union Catalog for records containing *lav* (Latvian) as the language fixed field code; *Latvia, Latvian*, and *Baltic* in fields 650 and 651 (Library of Congress (LC) subject headings); and selected sets from LC classification numbers PG and PZ.

### Findings

The extracted Latvian file consisted of 8,044 records, .05% of the total 15,801,330 records then in the Online Union Catalog. The great majority (98.5%) of records were for print or microform materials, of which 96% were monographs and 4% were serials. There were 27 languages represented in the Latvian file, of which Latvian (69.4%), Russian (10.6%), English (8.9%), and German (7%) were the most common. The other 23 languages combined represented only 4.1% of the Latvian file.

Cataloging for the 8,044 Latvian records was supplied by 329 institutions. This represented only 8.09% of active institutions, so not only was the total percentage of Latvian records small, but their contribution came from few institutions. The largest contribution was from LC tapeloads (1,646 records, 20.46%), followed by Kent State (Ohio) University (1,399 records, 17.39%) and Cleveland Public Library (1,104 records, 13.72%). These three institutions, which represented only .9% of source institutions, supplied 51.57% of Latvian records.

There were 35,030 holdings symbols in the Latvian file contributed by 1,796 institutions. No institution had over 2,000 holdings for Latvian records and only four—Kent State University, University of Wisconsin, Cleveland Public Library, and the Library of Congress—had between 1,001 and 2,000 records. Fewer than 5 records were held by each of 1,093 (60.85%) institutions and, of these, 481

(26.78%) held only one item. It was discovered that some of the institutions having the larger holdings of Latvian records actually had significantly more Latvian materials than were represented in the Online Union Catalog.

The analysis of holdings by library type showed that academic and academic research libraries (927; 51.61%) had the vast majority of holdings (24,946; 71.21%). The only other significant holdings were in public libraries (16.9%) and federal libraries (6%). This is accounted for primarily by the holdings of the Library of Congress.

The earliest Latvian records dated back to the seventeenth century; records predating the twentieth century formed a small group of titles (1.31%). Records dated 1941–1987 formed the largest part of the file (7,188; 89.35%); of these, 5,160 (92.4%) were in the Latvian language.

One of the original aims of this study was to establish the extensivity of the Latvian records set in comparison with what is available. It was possible to carry out this part of the study only for Latvian language records for items published in the West between 1940 and 1980; bibliographies for materials published in the Soviet Union were not available. There were 2,899 Western imprint Latvian language records in the file, and these represent about 41% of available Latvian language materials listed in the *Bibliography of Latvian Publications Published outside Latvia* (Jegers 1957).

Data for the analysis of subject distribution were obtained by searching fields 050 (LCCN) and 090 (locally assigned LCCN). Of the 8,044 Latvian records, 6,247 (77.66%) were found to have LCCNs. Subject coverage of the records was heavily concentrated in the areas of language/literature, history, and social sciences.

### Database Quality

Problems were identified in database quality of the Latvian records in the following areas: duplicates, language fixed field, descriptive cataloging and authority file control, and LCCNs. Duplicate records were observed while browsing through collective displays of voluminous Latvian authors. A small random sample of records, which were checked against each other manually and online, revealed a possible duplication rate of between 5% and 22%. The greatest problem faced by source institutions seemed to be a lack of catalogers with a knowledge of the Latvian language. This is exacerbated by lack of attention to details of the language, particularly the use of diacritics, when entering data and proofreading. In the language fixed field, some Latvian records (code *lav*) were coded *lat* (Latin) or *lit* (Lithuanian). This casts some doubt on the statistical language analysis mentioned earlier.

In all the descriptive cataloging fields, but particularly in personal and corporate name headings and titles, the most common problem was omission of diacritics or using the wrong keys to enter them. Each incorrectly entered diacritic causes the same name to appear multiple times when printed in an alphabetical list. Lack of knowledge of Latvian grammar also caused a number of personal name headings to be cataloged under Christian name rather than surname, and the genitive rather than the nominative case was used. Using both initials or expanded Christian names prevents collocation and efficient retrieval. Lack of authority control was also noticed for personal names with variations or compound surnames, resulting in name headings under all possible variations. Mistakes in field 260 (place of publication and publisher) are serious problems because this field is used to check for duplicate records.

Few LC classification numbers relate specifically to Latvia, but geographical subdivision of certain subjects is possible. In such cases, two problems were not applying the geographical subdivision where provided and applying them inconsistently. A number of examples taken from different LC schedules showed that LC itself has not consistently applied its own classification scheme.

### Conclusion

It was suggested that this study be considered as a pilot project for further research, perhaps in other European languages that are inflected and have diacritics. We discovered enough problems in the database to recommend that the incorrect data in the Latvian file be further identified and systematically repaired. It is also suggested that, because no work has been done on the quality of LC subject headings in the Latvian file, research in this area could be useful to supplement and further substantiate research already carried out by other sampling projects.

### References

Jegers, Benjamins. 1957. *Bibliography of Latvian Publications Published outside Latvia*. Vasingtona: Amerikas Latviesu Apvienibas Kulturas Fonds.

This research project was jointly funded by the British Library Research and Development Department, U.K., and the OCLC Office of Research.

# Mercury: An Electronic Library

**Project Team:** William Y. Arms, Mark H. Kibbey, Thomas J. Michalak, James H. Morris, Dana S. Scott, Marvin A. Sirbu, Carnegie Mellon University; Martin Dillon, Michael J. McGill, OCLC

**Abstract**
OCLC and Carnegie Mellon University are collaborating to develop and implement an electronic library allowing remote access to text and graphics stored on various media.

The vision of an electronic research library is simple: to deliver information needed for research to a scholar's workstation. To achieve this goal, OCLC and Carnegie Mellon University have joined forces to develop and distribute an electronic library known as Mercury. This effort has three primary goals:

- To demonstrate the feasibility of a large-scale distributed library developed using current technology
- To serve as a laboratory for studies in electronic information delivery
- To provide electronic library services to a community of scholars and researchers

### System Overview

When the system is operational, users will access the Mercury library using powerful desktop workstations well suited to the manipulation and display of text, mathematics, and graphics on large, high-resolution screens. These workstation computers will be connected to a network allowing access to remote databases and other computing resources. Information accessed by users will be stored on computers known as "servers." All elements of the system will be connected to a large, high-speed communications network so that, in effect, each component will be interconnected (figure 2.1).

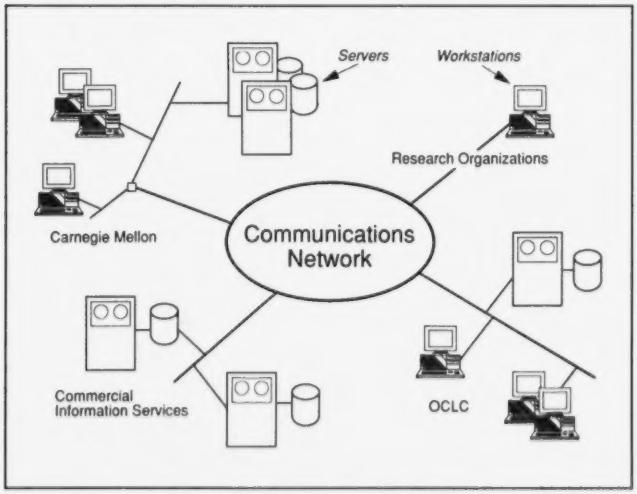



**Figure 2.1** System Overview

The Mercury system will be based on a distributed, campuswide computer network that was developed in partnership with IBM and is now installed on the Carnegie Mellon campus. The system has three major components: (1) communications network, (2) servers, and (3) high-function workstations.

Whenever possible, the Mercury system will use standard protocols to ensure compatibility with the International Standards Organization (ISO) Open System Interconnect (OSI) model. Protocols for communication, file transfer, remote call procedures, searching, database manipulation, and software applications among other functions are necessary to exchange information between different computers and different networks. Additionally, standard methods for tagging the logical components and structure of electronic text are needed. The Mercury system will serve as a test bed for the study of protocol standards and applications. The project will rely in part on information gained in the Linked Systems Project which was conducted by the Library of Congress, The Research Libraries Group, and OCLC.

### The Reference Assistant

Information available on the Mercury system will be stored on various media in different locations on the network (figure 2.2). Some databases may be local, resident on disk or stored on compact disc (CD). Such databases might include dictionaries and other reference works, journal articles or personally developed files. Other data will be stored in databases remote from the user's workstation. In this case, the user may need assistance to select the appropriate database for searching.

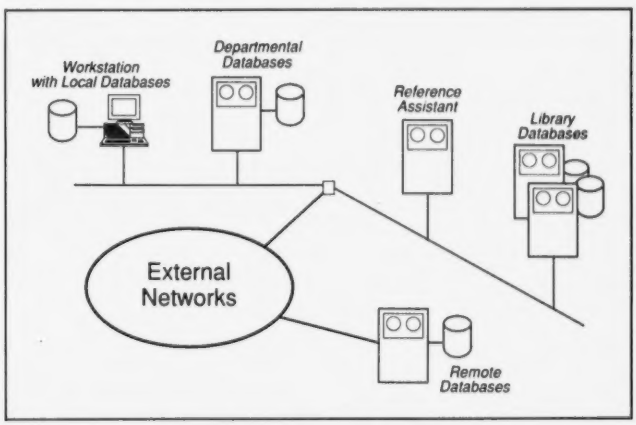



**Figure 2.2** Data Storage Locations

A server called the reference assistant will assist users to select databases and locate information. Development of the reference assistant will proceed in three stages. Initially, this server will only present the user with a list of database names. This will be augmented by a catalog of services or an index to available database services. Finally, the reference assistant will be an expert system that will interact with the user to determine the information need, develop

and refine search queries, select and connect to appropriate databases, and perhaps suggest alternative sources of information such as electronic bulletin boards or other users on the network.

### The Library

The Mercury system will initially serve as a large-scale electronic library laboratory used primarily by computer scientists. The Association for Computing Machinery (ACM) will provide machine-readable versions of the following journals: *Computing Reviews* (past 10 years), *Collected Algorithms* (past 25 years), *Communications* (past 2 years), and *Guide to Computing Literature* (past 10 years). In addition, the Institute of Electrical and Electronics Engineers (IEEE) has provided sample machine-readable files.

Unpublished electronic working papers and bulletin boards will provide another source of information, as will data obtained from networks such as ARPANET and commercial database providers. This will ensure sufficient machine-readable files for experimentation, analysis, and research.

### System Integration

The Mercury system will integrate components from various sources. The CMU network will be extended to include search servers. These computers will serve as test beds to monitor the performance of various searching algorithms. Initially, the search servers will provide free-text searching using a commercially available search system and will communicate with the reference assistant and workstations.

The search server will use conventional magnetic disk storage for data and indexes. Two other storage media will be used. An image server will store digitized information using write-once-read-many-times (WORM) optical discs. Typically, images will be scanned, digitized, and stored on disc. The image server will be a test bed for research on methods of storing very large raster images which will be used in conjunction with textual information retrieved from the search server.

Other data could be stored locally on CD. While useful as a dense storage medium for local databases, CDs provide relatively slow access to information and are generally not suitable for use as search servers.

Free-text search systems require indexing. This demand on computational resources often degrades system performance. To alleviate this problem, the Mercury system will use a separate index builder to generate indexes as new information enters the system.

### Conclusion

The goal of this ambitious project is to demonstrate that the electronic library of the future is feasible today using current technology. Implementing the system will test accepted and emerging protocols for intersystem communications, remote call procedures, and sophisticated database searching on multiple databases using various storage media. The system will also function as an electronic library to support research in computer science by delivering information to the scholar's desktop workstation.

## Term Weighting in Document Retrieval

**Principal Investigators:** Clement T. Yu, Professor, Department of Electrical Engineering and Computer Science, University of Illinois at Chicago; Hirotaka Mizuno, Engineer, OKI Electric Industry Production Engineering Center, Tokyo

**Abstract**

Two methods are given to improve weighting schemes by using relevance information from a set of queries. The first method estimates parameter values of two independence models in information retrieval: the binary independence model and the nonbinary independence model. The estimated parameters are used to calculate optimal weights for terms in a different set of queries. Performance of this estimation is compared with the inverse document frequency method, the cosine measure, and the statistical similarity measure. The second method is to learn optimal weights of the nonbinary independence model adaptively by a learning formula. For both methods, experiments were performed on three different document collections, CISI, MEDLARS, and CRN4NUL, and results are reported. Both methods show improvements compared with existing weighting schemes. Experimental results show that the second method gives slightly better performance than the first, and has simpler implementation.

In information retrieval systems, it is common to compute the similarity between a query and a document for retrieval purposes. Many methods of calculating similarities have been presented. Two well-known methods are the inverse document frequency method (Jones 1972) and the cosine measure (Salton and McGill 1983). The inverse document frequency method assigns a higher weight to a term that occurs in few documents than a term in many documents. The similarity measure between a document and a query is simply the sum of the weights of the query terms in the document. The cosine measure is the cosine of the angle between the query and the document when they are represented as vectors. In spite of their intuitive appeal, they cannot ensure optimal retrieval.

The binary independence model (Yu and Salton 1976; Robertson and Sparck 1976) and the nonbinary independence model (Yu and Lee 1986) are theoretical models that have been proposed to guarantee optimal retrieval. Optimality here means that a document with higher probability of relevance to a query is assigned a higher similarity value than a document with lower probability of relevance.

These two models are theoretical because they suppose a priori knowledge about some parameters used to calculate similarities.

In this report the parameter values of these two independence models are estimated from document and term frequencies. The performance of the estimated independence models are compared with those of the inverse document frequency method, the cosine measure, and the recently reported statistical similarity measure (Wong and Yao 1987). A learning scheme is given. Both the estimated nonbinary independence model and the learning scheme give better performance than existing indexing methods.

### Binary Independence Model

According to the binary independence model (Robertson and Sparck; Yu and Salton), the weight assigned to term $T_j$ is given by

$$w_1(T_j) = \log \frac{p_j}{1-p_j} - \log \frac{q_j}{1-q_j} \qquad (1)$$

where $p_j$ is the probability that a relevant document has the term $T_j$ and $q_j$ is the probability that an irrelevant document has the term $T_j$.

In deriving the weighting function (1), we suppose a priori knowledge of $p_j$ and $q_j$. These probability values are unknown in advance under ordinary circumstances. Our aim is to estimate $p_j$ and $q_j$ from some other easily obtainable parameters.

### Nonbinary Independence Model

In the nonbinary independence model, the independence assumption made previously is retained while the binary assumption is discarded. In other words, we take into account the term frequency of each term in each document. Then the optimal weight of a term $T_j$ with frequency of occurrences $d_{ij}$ in a document $D_i$ is calculated by the following formula (Yu and Lee)

$$w_2(T_j, d_{ij}) = \log \frac{p(tf=d_{ij}|R)}{p(tf=d_{ij}|I)} - \log \frac{p(tf=0|R)}{p(tf=0|I)} \qquad (2)$$

where $p(tf=d_{ij}|R)$ is the probability that a relevant document has $d_{ij}$ occurrences of the term $T_j$, and $p(tf=d_{ij}|I)$ represents the probability that an irrelevant document has $d_{ij}$ occurrences of the term $T_j$. $R$ and $I$ stand for *relevance* and *irrelevance*, respectively; $p(tf=0|R)$ and $p(tf=0|I)$ are the probabilities with no occurrence of the term.

Again, the probabilities $p(tf=d_{ij}|R)$ and $p(tf=d_{ij}|I)$ are usually unknown in advance. In addition to the estimation of $p_j$ and $q_j$, the probabilities $p(tf=d_{ij}|R)$ and $p(tf=d_{ij}|I)$ are estimated. The probability values of the two independence models described in (1) and (2) are estimated by linear regression (Yu and Mizuno, 1987).

### Experiments

Three different document collections are used to evaluate the performance of the two estimated independence models. Experimental results show that the estimated nonbinary model gives better retrieval performance than the inverse document frequency method, the cosine method, and the statistical similarity measure in each of the three collections.

**Learning scheme.** A learning formula used in the experiments is as follows:

$$w_{k+1}(T_j,tf) = w_k(T_j,tf) + \frac{c}{n_j}[w_{opt}(T_j,tf) - w_k(T_j,tf)] \qquad (3)$$

where $w_k(T_j,tf)$ is a weight of $j^{th}$ term $T_j$ with term frequency $tf$ after learning $k$ times, $w_{opt}(T_j,tf)$ is optimal weight of $j^{th}$ term $T_j$ with term frequency $tf$ by the nonbinary independence model, $c$ is a learning coefficient, and $n_j$ is document frequency of term $T_j$. Although $w_{opt}(T_j,tf)$ is determined not only by $T_j$ and $tf$ but also by a query in a learning query set, we do not express it explicitly here for simplicity of expression; $w_k(T_j,tf)$ and $w_{opt}(T_j,tf)$ are abbreviated as $w_k$ and $w_{opt}$ in the sequel.

The reason document frequency $n_j$ is incorporated in this formula is that weights of term $T_j$ are learned $m \times n_j$ times at every iteration of learning, where $m$ is the number of occurrences of term $T_j$ in the learning query set. This means that weights of high document-frequency terms are learned many more times than weights of low document-frequency terms. Document frequency $n_j$ in the formula (3) prevents this inequality in learning.

### Experimental Results of the Second Method

Experiments were performed on three different document collections. The query set for learning ($Q_1$) and evaluation ($Q_2$) were disjointed and are as those in the previous experiments. Five different learning coefficients ranging from .002 to .1 were tested, and performance was measured in comparison with the inverse document frequency method. The results are summarized as follows:

1. Improvement in performance is observed in every collection when the learning coefficient is small (.002–.016).
2. Improvement of MEDLARS is small compared with CISI and CRN4NUL.
3. Performance rapidly deteriorates in CRN4NUL after the first learning with the coefficient .1.

The small improvement in MEDLARS may be attributed to the size of learning data. In MEDLARS, 20 queries are used for learning in contrast to 61 queries in CISI and 110 queries in CRN4NUL. Additionally, a term in the query evaluation set $Q_2$ does not always appear in the query learning set $Q_1$. Only 22 out of 88 query terms contained in $Q_2$ appear in $Q_1$.

A weight of a query term of every term frequency approaches its optimal weight after each learning time when a learning coefficient is small. On the other hand, when a learning coefficient is large, a weight of some query terms may be overestimated.

## References


Jones, Karen Sparck. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (March):11–20.

Robertson, S. E., and Karen Sparck Jones . 1976. Relevance weighting of search term. *Journal of the American Society for Information Science* 27 (May–June):129–46.

Salton, Gerard, and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Wong, S. K. M., and Y. Y. Yao. 1987. A statistical similarity measure. In *Proceedings of the Tenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, New Orleans, June 3–5, 1987*, eds. C. T. Yu and C. J. Van Rijsbergen, 3–12. New York: ACM Press.

Yu, Clement T., and T. C. Lee. 1986. Non-binary independence model. In *The 1986 ACM Conference on Research and Development in Information Retrieval, Pisa, Italy, September 8–10, 1986*, ed. Fausto Rabitti, 265–68. Pisa: Presso de Officine Graphiche della Lito-Tipografia.

Yu, Clement T., and M. H. Mizuno. 1987. Two learning schemes in information retrieval. Technical report. University of Illinois at Chicago: Department of Electrical Engineering and Computer Science.

Yu, Clement T., and Gerard Salton. 1976. Precision weighting: an effective automatic indexing method. *Association for Computing Machinery Journal* 23:76–88.

# 3

# Library and Information Science Research Grant Program

In January 1988 the OCLC Office of Research initiated the Library and Information Science Research Grant (LISRG) program. The LISRG program, which replaces the Library School Research Equipment Support (LSRES) program, assists schools of library and information science to conduct high-quality technical research.

This grant program funds release time from teaching for the principal investigator, research assistants, travel, equipment, and other project-related expenditures. Grants are limited to $10,000. The projects are expected to take no longer than one year from the date of the award. Upon completion final reports are submitted to OCLC; research results must be nonproprietary and in the public domain.

Participants are required to submit a technical proposal that (1) describes the nature, scope, and method of the proposed research, (2) explains why the research is innovative through references to related research reported in the literature, (3) proposes the significance of research to library and information science, (4) explains how requested equipment is important to the research, (5) identifies methods that will be used to measure the success of the project, and (6) suggests future research that might build on the anticipated results of the project. ►

Each proposal is submitted to a two-tiered review process. Initially, the proposal is distributed to two OCLC staff for review. These reviews and the proposals are then submitted to a panel for further evaluation. The panel, representing various perspectives, meets with the director of the Office of Research to make funding recommendations. This year's panel members are Michael Kathman, Director of Libraries, Media, and Academic Computing, St. John's University/College of St. Benedict; Mary Ann Mercante, Executive Director, Missouri Library Network Corporation; Elaine Svenonius, OCLC Visiting Scholar; and Richard Van Orden, Manager, EIDOS Information, OCLC.

The 1988 grant recipients and preliminary reports from the 1987 Library School Research Equipment Support program follow.

**Library and Information Science Research Grant Awards**
**June 1988**

| | |
|---|---|
| Title: | Research of Knowledge-based Descriptive Cataloging of Cartographic Publications: An Experimental Advice-giving System, Mapper. |
| Principal Investigator: | Harold Borko, Professor |
| Institution: | University of California, Los Angeles |
| Award: | $7,055 |
| Title: | An Automatic Method of Enhancing Topical Searching for Online Catalogs Based on Classification Clustering |
| Principal Investigator: | Ray R. Larson, Assistant Professor |
| Institution: | University of California, Berkeley |
| Award: | $10,000 |
| Title: | The Relationship between Library Holdings of Small Publishers' Books and Selection Sources, Particularly Reviews |
| Principal Investigator: | Judith Serebnick, Associate Professor |
| Institution: | Indiana University |
| Award: | $2,000 plus an M310 Workstation and printer |

# Analytical Study of Bibliographic Data on a Title Page

**Principal Investigator:** Billie Grace Herring, Associate Professor, Graduate School of Library and Information Science, University of Texas at Austin

**Abstract**
The objective of this study is to analyze bibliographic data on the title page of a modern monograph in the English language as a natural language artifact. Bibliographic data are examined in terms of visual and linguistic characteristics. The intent is to contribute to an understanding of the title page on its own merits as a specialized form of natural language. Such knowledge may then assist in understanding the title page as a kind of structure for which some decision rules can be developed in an expert system for descriptive cataloging.

Title pages have long been the source of cataloging data. What is there about the language of the title page that enables catalogers to understand and interpret the title page? This study examines the bibliographic data on a title page as a natural language artifact and the title page as a subset of natural language which has both syntactic and semantic characteristics. The assumption is made that catalogers use a set of informal rules to understand and interpret title pages, and that these rules are similar to those used to understand natural language. To investigate the possibility of developing a set of rules applicable to the future construction of expert systems for descriptive cataloging, this study examines the visual and linguistic characteristics of title page data to identify patterns that are consistent across the sample.

## Objectives

The overall objective is to study the modern title page as a phenomenon and to analyze bibliographic data on the title page as a natural language artifact. More specifically, the role of the title page in descriptive cataloging is examined by identifying to what degree bibliographic data on the title page are included in the bibliographic records found in

the OCLC Online Union Catalog. Additional objectives are to identify the following: (1) visual characteristics of blocks of words and separators among bibliographic elements on title pages, (2) categories and functions of words and blocks of words, (3) syntactic and semantic roles of the placement of blocks of words, (4) semantic categories and functions for blocks of words, and (5) potential applications of the syntactic and semantic characteristics of title pages to the construction of knowledge bases for descriptive cataloging.

### Assumptions and Limitations

Current title pages are very different from pretwentieth century title pages; thus, only books published since 1900 are included in the sample. The title page is examined only as the source of information used in cataloging practice, not as a product of the book design and publishing process or as the basis for descriptive bibliography. It is assumed that the semantic and syntactic characteristics examined reside within the language of the title page and are independent of a cataloger's perception of the title page. It also is assumed that data on the title page of monographs are presented differently from that of other formats such as serials, audiovisual materials, or technical reports. Also, data from the title page verso is studied as a part of the source of data for the bibliographic description, but is not studied in the syntactic and semantic analysis.

### Methodology

A sample of 200 books was randomly selected from the Perry-Castenada Library of the University of Texas at Austin, and the title page and verso of each book was photocopied. The Online Union Catalog was searched to find the identifying bibliographic labels for data on the title page. The bibliographic data were grouped into blocks of word strings separated by visual boundaries. Each block was labeled according to the bibliographic element it represents.

The visual attributes of each block are a group of primitives indicated by position of the block on the title page, portion of title page occupied by a specific block, relative type size of data in the block, font, horizontal alignment, and darkness or boldness of characters. Other visual elements examined are the coordinates or separations indicated by graphic symbols, ruler lines, punctuation marks signifying the end of blocks, use of space between blocks, and rotation of blocks to vertical. Visual attribute data are being analyzed using descriptive statistical procedures.

Blocks of words were grouped into nine categories according to the bibliographic elements they contain. Lexical analysis of this data is just beginning. Tokens found in the bibliographic elements have been grouped. Word tokens have been further subdivided and analyzed in terms of part of speech and frequency in each bibliographic area. Proper names were examined in terms of morphological rules and were classified into further subcategories. Phrase patterns were labeled and identified according to the patterns they display within each bibliographic element.

### Preliminary Findings

Preliminary analysis indicates consistency with a related study of data (Jeng 1987) identified with visual and linguistic elements and, to some extent, with that reported by Weibel (1987). Because the sample size is small, conclusions will, of necessity, be guarded and tentative. It does appear, however, that sufficient data about patterns exist and that a few tentative decision rules for visual coordination can be proposed.

### References

Jeng, Ling Hwey. 1987. The title page as the source of information for bibliographic description: an analysis of its visual and linguistic characteristics. Ph.D. diss., The University of Texas at Austin.

Weibel, Stuart L. 1987. Automated title page cataloging. In *Annual Review of OCLC Research, July 1986–June 1987*, 3–4. Dublin, Ohio: OCLC Online Computer Library Center, Inc.

## Applying the Revised MARC Format for Three-dimensional Objects in Museums

**Principal Investigator:** Esther G. Bierbaum, Associate Professor, School of Library and Information Science, The University of Iowa, Iowa City

**Abstract**

This study compares the data requirements of museum registration (accession and description) and cataloging with the data fields provided in the revised MARC format for visual materials. Museum records are not standardized. A consensus model for records was based on museum literature and consultations with museum registrars. A work form was also developed. The data requirements of small museums appear to be met in the new MARC format. When the changes are fully implemented on the OCLC Online System, museum registrars will be able to evaluate several possible record formats.

This project examines the suitability of the revised MARC format for visual materials for creating records for three-dimensional (3-D) objects in small museums where there is the greatest need for assistance in registering objects (that is, making records of accession and description) and cataloging collections and where paper systems still prevail.

Museums and libraries share collection control problems such as creating surrogate records of items in the collection to facilitate access and to compose a catalog of the collection. Museum methodologies and formats vary greatly; there is no equivalent of the *Anglo-American Cat-*

*aloguing Rules*, 2d. ed. (*AACR2*). In fact, before *AACR2*, libraries also depended largely on "local usage" for nonbook cataloging. Although *AACR2* provides a standardized and uniform bibliographic description of three-dimensional objects, the next step—creating a unique MARC format for this medium—was not taken. Instead, the description of objects was subsumed in the format for audiovisual materials. The Library of Congress issued a revised MARC format for visual materials in January 1987. Using this format, museums could presumably use an online bibliographic network database to generate and store bibliographic and collections records.

### Museum Records

The first phase of this project has been the identification of essential data fields in the descriptive record of an object. Individualized, nonstandardized records are the norm for museums. Each institution constructs its registration system and organizes its catalogs according to local needs, collections, and personnel. Automation, when introduced to museums, tends to emphasize report generation at the system level, largely overlooking consistency in form and content at the record level. Because sharing records is not common practice, there is little incentive for uniformity in cards or online records.

There is general agreement in the literature and in practice that the following data comprise the minimum information content in the registration record: (1) control number (usually registration number); (2) date, source, and mode or method of acquisition; (3) name(s) of object (scientific, generic, functional, or common); (4) description of object, including materials and dimensions; (5) history/provenance of object; and (6) location within collection.

Basing fields and content on the minimum record information, the investigator developed a model data entry format. This format was then compared with the revised MARC format for visual materials to produce a proposed minimal-level MARC entry work form for museum objects and a supplementary schedule of extended data fields. The resultant format is similar to the standard MARC format for visual materials and presented no difficulties to an advanced cataloging class when used as a pencil-and-paper exercise.

A set of decision rules was also provided in the form of "if/then" and "when/do"; one expects that such formulaic statements will be useful to museum registrars. Testing using keyboards was deferred because of the delayed arrival of equipment and because the revised MARC format had not yet been fully implemented by OCLC.

The following fields should accommodate the essential data content:

| Field | Description |
|---|---|
| 033 | Date/place of capture/finding |
| 035 | Control number (e.g., registration) |
| 1xx | Name, in the case of a creator, or |
| 245 | Name, as title proper, also in conjunction with subfield ǂh. |
| 260 | Place and agent of creation/manufacture |
| 300, 340 | Physical description |
| 541, 561 | Acquisition and provenance notes |

In addition, the following fields which are new, have expanded meaning or increased utility for recording objects would be functional in all but the most minimal records:

| Field | Change or Comment |
|---|---|
| Leader/06 (Type) | Expanded definition of "r" |
| 008/33 (Type mat) | Addition of "c" (art reproduction) and "w" (toy) |
| 250 (Edition statement) | Provision for nonedition statements, e.g., version |
| 5xx (Notes) | Other notes areas are useful |
| 7xx (Added entries) | As necessary for collocation |
| 754 (Taxonomic I.D.) | Important in certain collections |
| 773 (Host item) | Field/ground indexing (cf. 340) |

While these fields do not exhaust the possibilities of the revised MARC format for visual materials, they fulfill stipulations for important data fields found in the literature and cited by museum registrars. The limited collection might not use fields 340 or 773; the larger collection might add other notes and create more added entries and access points.

The new fields and subfields are not yet available on the OCLC Online System, although the expectation is for an early completion of the software conversion. OCLC has indicated that it will not print or index fields 033 and 541; it will print, but not index, fields 561 and 773. Additionally, fields 340 and 754 are not included in the second edition of *Audiovisual Media Format*. Until the format is fully implemented, it is difficult to judge the effects of these variances. A field of potential value to 3-D collections is OCLC's non-LC field 910, user-option data. Printing is confined to 21 characters, but up to 1,230 may be retained in the archive record. Field 910 could be used for location, subcollection, and similar notations.

### Record Access

Faced with the problem of creating collocation points for the object records, museums have customarily reordered the data in the record, filing the record according to the newly created access point. Accordingly, the same object has different records for different retrieval purposes in the donor file, the provenance catalog, the exhibit location file, and so on.

Thus, one difficult problem associated with adapting library-style unit entry records to museum files is determining what type and means of access to provide. In an online system, the record may be downloaded and the data manipulated by the local system to produce indexable retrieval modes; in a card system, the points of access must be so identified that the network system will produce cards corresponding to the selected points in form, format, and content. This highlights the importance of tracings and the headings resulting from them. The next phase of this project will test variances in added entries for suitability in a museum catalog.

### Further Testing

At this writing, the author is preparing test sets of added entries for evaluation by a curator at the Iowa Historical Society. One method uses 7xx added entry data from notes and other fields; another uses thesaurus-based subject headings in the 6xx fields. The test should indicate how decisions regarding record profile may be made by the local agency. Testing of the overall construction and input of museum registration records can go forward when the revised format is completely installed on the database.

The full revised MARC format for visual materials has much to recommend it from the librarian's viewpoint. As it becomes more widely known, the input of records for 3-D materials in libraries should increase. At the same time, the various forces in the museum environment that seem to be guiding those agencies toward resource sharing may encourage a standardization of records so that collection information can be exchanged. Whether the small collection can adapt registration records to the OCLC version of the MARC format for visual materials cannot yet be ascertained. If such input proves feasible, museums will benefit from the ability to create and share both bibliographic and object records in a network environment.

### References

*Anglo-American Cataloguing Rules*. 2d ed. 1978. M. Gorman and P. W. Winkler, eds. Chicago: American Library Association.

*Audiovisual Media Format*. 2d ed. 1986. Dublin, Ohio: OCLC Online Computer Library Center, Inc.

Library of Congress. 1987. *MARC Formats for Bibliographic Data, Update No. 15*. Washington, D.C.: Library of Congress.

Light, R. B. et al., eds. 1986. *Museum Documentation Systems*. London: Butterworth's.

# Converting English Subject Headings into Spanish

**Principal Investigator:** Emilia Bernal-Rosa, Professor, Graduate School of Librarianship, University of Puerto Rico—Rio Piedras

**Abstract**

This study attempts to provide multinational access to subjects by converting English subject headings into Spanish. A pilot bilingual English-Spanish controlled vocabulary is being developed. The pilot vocabulary will be tested and analyzed considering the needs of non-English speaking catalogers and subjects searchers.

The OCLC Online Union Catalog, a major international cataloging and bibliographic database, is prepared for English-speaking users, and subject headings in OCLC bibliographic records are in English. Subject headings in a dictionary catalog for non-English-speaking users must be translated. Possible solutions to the language problem include bilingual or multilingual controlled vocabularies through which the subject headings could be changed by a cataloger into their foreign language equivalents.

### Scope and Methodology

This study is creating a pilot bilingual English-Spanish controlled vocabulary and applying it to cataloging and subject searches using the Online Union Catalog as a resource. (Subject searches are conducted using EASI Reference™, a database of OCLC bibliographic records available through BRS® Information Technologies, Inc.)

We sought the advice of library users in Spanish-speaking countries in the Caribbean and Latin America, where the dictionary catalog is commonly used, to help determine the criteria for the pilot vocabulary. The sample English vocabulary from which the bilingual vocabulary will be built is being chosen from a Spanish language subset of the Online Union Catalog. A systematic sample of subject headings is being selected from topical subject headings in OCLC bibliographic records for monographic materials in the social sciences (Dewey Decimal classification 300, excluding 390, and Library of Congress (LC) classification H-L) cataloged during the last ten years.

After this process is completed, the Spanish equivalents to the subject headings will be identified, thus creating a bilingual vocabulary. The Engish-Spanish sections of three subject headings lists will be used to determine the equivalents.[1]

### Vocabulary Analysis and Testing

Once the bilingual controlled vocabulary is prepared, it will be subjected to (1) a quantitative and comparative analysis of the different components of the list, and (2) performance tests to determine how well the list will serve as a bilingual cataloging and searching tool for the Online Union Catalog.

For English subject headings taken from the Online Union Catalog the analysis will include (1) the number of subject headings per record; (2) type of subject headings (single, compound, with subheadings) and frequency by type; (3) types of subheadings (form, geographic, topical) and frequency by type; (4) proportion of LC determined subject headings vs. non-LC determined subject headings; and (5) degree of deviation of the subject headings from those in *Library of Congress Subject Headings*, 10th ed., and the most recently updated LC list.

For the Spanish section of the bilingual list analysis will determine how well each list represents the subject matters identified through the OCLC database. Results obtained from each list will be compared with the combined vocabulary to determine statistical significance and to contrast in each list the degree of correspondence in meaning between the Spanish terms and their English equivalents for LC and non-LC determined subject headings.

As a performance test, an exercise on assigning subject headings in Spanish will be executed by groups of library school students who can work with English and Spanish subject headings. The Online Union Catalog will be used as primary cataloging source in the exercise and the subject headings will be converted to Spanish using the bilingual controlled vocabulary. Evaluation of performance for each list in Spanish and for a combined list of Spanish terms will be based on weighted degree of accuracy of the subject headings in Spanish in relation to English subject headings in the database records. Individual scores will be added to produce group scores, and differences in scores will be tested for statistical significance. Average time taken by the groups to complete the assignments will be measured. An exercise in searching the OCLC database starting with topics in Spanish will be prepared to identify materials that can be used to enrich a collection in specific subject areas. Subjects will be given in Spanish, translated into English with the vocabulary, and used to identify pertinent material in the Online Union Catalog. Evaluation will be in terms of the following:

1. Relevance of the material selected using each Spanish list and a combined list as expressed in the formula

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of documents retrieved}}$$

2. Average time taken using each list to complete the assignment
3. Relationship of (1) to (2)

By combining *BILINDEX*, the *Lista de Encabezamientos*, and a local Puerto Rican list, a new and better tool is prepared and one that could be applied with versatility at the local and regional level.

### Notes

1. California Spanish Language Data Base. 1984. *Bilindex: A Bilingual Spanish-English Subject Heading List*. (Oakland, Calif.: California Spanish Language Database).

Organization of American States. 1985. *Lista de Encabezamientos de Materia para Bibliotecas*, 2a ed. (Bogota: Procultura).

Colegio Universitario de Humacao, U. P. R. Biblioteca. 1984. *Informe Final de Epigrafes*. (Rio Piedras, Puerto Rico: Colegio Universitario de Humacao, U. P. R. Biblioteca).

# Evaluating Subject Collections

**Principal Investigator:** Howard D. White, Professor, College of Information Studies, Drexel University, Philadelphia

**Abstract**

A new way of setting or verifying collection levels makes use of brief checklists chosen by a subject expert and scaled with OCLC holdings data. A trial evaluation is performed using two subsets chosen by subject experts. In preliminary trials at two large research libraries, the test appears to discriminate well.

Using OCLC data on how frequently titles are held by libraries, I have developed a brief test that assigns libraries a score for existing collection strength in a subject area. The test can assist in setting or verifying collection intensity levels, which are now set subjectively and verified only with great labor. Results for tests in two subject areas administered at two large libraries will be presented.

### Background

Led by the Research Libraries Group (RLG), librarians nationwide are stating publicly at what level they collect materials in major Library of Congress (LC) or Dewey subject categories. Usually the statements are part of written policies such as the *RLG Conspectus* (RLG 1985). The levels, also called collection intensity indicators, make up an ordinal scale: 0, out of scope; 1, minimal; 2, basic information; 3, instructional support; 4, research; 5, comprehensive. The scale represents librarians' judgments of past (existing collection strength) and planned (current collection intensity) subject development activity. The *Conspectus* calls for judgments in hundreds of subject areas defined by LC class ranges.

To gain insight into how accurately librarians estimate their collections, RLG and other academic libraries have sponsored verification studies. Participants check their holdings against specially prepared lists of up to 1,000 items, and RLG matches the coverage scores against previously claimed ratings. Libraries with very different scores are found to have claimed equal ratings, often a 4. Sometimes scores suggest ratings should actually be reversed. A library claiming a level 3, for example, may have better coverage than one claiming a level 4.

On the checklists I have seen, moreover, the items chosen do not reflect differences in the intensity levels. There is no attempt to differentiate reference works (levels 1 and 2), items supporting undergraduate instruction (level

3), or specialized sources (levels 4 and 5). Instead, systematic random or judgment samples are taken from comprehensive scholarly bibliographies, and almost all items reflect the higher levels.

It should be noted, too, that collection intensity scales are cumulative in nature. A level 3 collection, for example, presupposes extensive holdings at levels 2 and 1. The checklists do not test this directly. It is assumed that if a library has extensive holdings of level 4 items, then its holdings at the lower levels are also extensive. At present, a test may suggest that a library should claim a level 2 because it has a relatively low percentage of items suitable for level 4. This seems a needlessly indirect way of verifying a level and raises questions about the validity of the test.

To do a typical verification study, then, a list of perhaps 1,000 items must be compiled—a major undertaking in itself—for participants to check against holdings. Whether done manually or by computer, the latter task is formidable and produces results of uncertain validity. Not surprisingly, RLG and other academic libraries have not done many verification studies. Yet there is a need for them in every LC class range, including many not yet tested. There is also a need for repetition of studies within a class if estimates are to be believed.

### A Briefer Test

In contrast to the checklist tests just described, I propose short tests of perhaps 40 items per subject category selected by a subject specialist. These brief tests can be compiled in a day or two and completed by one person in a few hours, particularly if the checking is done online.

The test distinguishes among libraries at levels 1 through 4. (Level 0, at which the library collects nothing, and level 5, at which the library attempts to collect everything, are not problematical.) RLG has stated that raters of collections should be particularly concerned with distinguishing level 3, which supports local needs, from level 4, at which point the collection becomes a national resource.

There are about ten items to check per level. These are scaled with OCLC holdings data to reflect both the definition of the level and the cumulative nature of the levels. Theoretically, holdings data should have a strong, inverse relationship with the collection-intensity scale. Monographs at the lowest level (1, minimal) should be held by many libraries, and monographs at the highest level (4, research) should be held by relatively few libraries. The other two levels are intermediate.

Since progressively fewer libraries should hold items as the levels increase, the levels form a scale of relative specialization, i.e., the degree to which titles appeal only to scholars rather than beginners or the general reader. For example, materials at levels 4 and 5 could be advanced, localized, foreign, academic, obscure or old. However, the notion of specialization is also akin to the notion of "difficulty" of items in the sense of "increasingly less popular, less likely to be acquired." A holdings count for each title thus measures its probability of having been acquired.

In making the test, titles are ranked low to high by their counts and then divided into quartiles. The ten items least frequently held become level 4 test items; the next ten, level 3; and so on, down to the ten most common at level 1. In "power" tests of this sort, items become increasingly difficult as the levels rise.

If the scale is cumulative, an exemplary library should have a coverage pattern that is strong at all levels. More particularly, if the library exceeds a threshold at a given level (e.g., 50% coverage), it should do at least as well at all lower levels. Because levels are scaled to be increasingly difficult to attain, a pattern of diminishing "percent covered" should emerge (table 3.1).

**Table 3.1** Sample Pattern

| Level | No. of Titles | (%) Covered |
|---|---|---|
| 1 | 10 | 100 |
| 2 | 10 | 90 |
| 3 | 10 | 80 |
| 4 | 10 | 30 |

### Methodology

Lists of about 40 titles have been solicited from subject experts. Each selector named about 10 titles per level. Tests based on lists in several subjects will be tried. This preliminary report shows results for library and information science (LIS) and nineteenth century Russian literature.

Holdings counts for titles on the two lists were retrieved from the OCLC Online Union Catalog. When works had closely related editions, counts for all non-microform editions were totaled and used. Titles were then ranked by their holdings counts and divided into quartiles, levels 1 to 4. Originally I planned to use only titles that were both (1) named by the selector to illustrate a given level, and (2) placed in that level when the holdings counts were ranked. However, selectors' assignments to levels differ about half the time from those produced by OCLC holdings counts. I let the latter determine levels (overriding selectors' judgments), but the selectors' attention to levels assures a good range of holdings counts.

The test lists were matched against holdings at Temple University and the University of Pennsylvania. Both university libraries have online catalogs covering materials cataloged since ca. 1972. The online catalog greatly facilitated checking. Card catalogs were also checked, however, and some items were found this way.

The selector for library and information science materials presented 40 items by title or author/title only. They were therefore treated as works: if any edition was held, the library was given credit for holding that title.

The selector for nineteenth century Russian literature presented 45 titles in MARC format. About one-third were in Russian. A few of the monographs, works of history or criticism in English, were credited regardless of edition held, but if edition seemed important for literary reasons, an exact match was required for credit. (English translations could not be substituted for Russian language editions; particular English translations of *Anna Karenina* and *Dead Souls* were required.) Serials on both lists were credited if the university had a current subscription.

The library and information science test list was selected by a librarian at Drexel University; the test for Russian literature, by a librarian at the New York Public Library. For obvious reasons, the tests were conducted on holdings at other institutions, Temple University and the University of Pennsylvania. It was important to learn whether those institutions could be distinguished in library and information science from nearby Drexel University, which presumably has a level 4 collection to support its library school. (This would be the first trial of the test's ability to distinguish level 4 from lower levels.) It was also important to test collections in Russian literature capable of reaching the higher levels.

### Results

While the test lists are too long to include here, the way the holdings counts structure them is interesting to see (and will be reported later). The LIS test was the less difficult of the two. For example, it contained some items that librarians might buy for their own purposes, regardless of curriculum and research at their schools (e.g., Sheehy's *Guide*, Gorman's *Concise AACR2*). Even if the LIS test had been harder, nineteenth century Russian literature seems intrinsically more demanding because of the linguistic and historical knowledge it requires. (Sorted by holdings counts, almost all titles at level 4 and half at level 3 were in Russian.) Nevertheless, the LIS test did not turn out to be too easy, nor the Russian literature test too hard. Both tests discriminated well.

Results generally confirm the major hypothesis. Percentages for LIS show a clear falloff as the levels rise, as predicted (table 3.2). If one adopts the convention that at least 50% coverage is needed to claim a level, both Temple University and the University of Pennsylvania are level 3 (which accords with my subjective impression).

The pattern for Russian literature (table 3.3) is irregular at level 3; Temple University has fewer titles than expected and the University of Pennsylvania has more. Again, using the criterion of at least 50% coverage to claim a level, the University of Pennsylvania emerges as a very solid level 3; Temple University, as a level 2 in this subject. In the *Conspectus*, for "Russian Literature—Nineteenth Century to Revolution" the University of Pennsylvania claims a 3 for both existing collection strength and current collection intensity. In the same two categories, Temple University claims a 3 (selected foreign-language material included with English) and a 2 (material primarily in one foreign language). The brief test corroborates these self-ratings.

**Table 3.2** Library and Information Science

| | | (%) Covered | |
|---|---|---|---|
| Level | No. of Titles | Temple University | University of Pennsylvania |
| 1 | 10 | 90 | 90 |
| 2 | 10 | 70 | 80 |
| 3 | 10 | 70 | 50 |
| 4 | 10 | 40 | 40 |

**Table 3.3** Nineteenth Century Russian Literature

| | | (%) Covered | |
|---|---|---|---|
| Level | No. of Titles | Temple University | University of Pennsylvania |
| 1 | 12 | 92 | 92 |
| 2 | 11 | 54 | 73 |
| 3 | 11 | 9 | 82 |
| 4 | 11 | 36 | 36 |

More testing is required to verify the present methodology. I leave that to a later report, content for now to argue that this sort of testing is more feasible, in terms of time and effort, than the large-scale sampling of the past. Ten items per level may be too few for stable percentages. The total of 40 was set so as not to overburden selectors and could be increased. But even 40 titles give useful results when structured as here.

### References

Research Libraries Group. 1985. *RLG Conspectus*. 1st vols.

# Measurement of Subject Scatter in the Superintendent of Documents Classification

**Principal Investigator:** Lee Shiflett, School of Library and Information Science, Louisiana State University, Baton Rouge

**Abstract**

Library of Congress MARC records containing Superintendent of Documents Classification (SuDoc) numbers for the Department of the Army, the Department of Agriculture, the Department of State, the Library of Congress, and a broad general sample were sorted by both schemes and the resulting orders compared. A new measure, normalized average minimum document movement, was proposed and computed, as were existing measures of dispersion resulting from reclassification. Findings do not support the hypothesis that the amount of dispersion of documents occurring when a collection is reclassified from SuDoc to Library of Congress Classification (LCC) is insignificant.

Government documents are normally arranged as a separate collection using the Superintendent of Documents Classification system (SuDoc) or are integrated with the main collection using the Library of Congress Classification system (LCC). The question of integration is reviewed by Waldo (1977) and Larson (1979). Would a documents collection arranged by LC classification have benefits that do not accrue from one organized by SuDoc?

The idea of a separate documents collection organized by LCC is certainly innovative. It is, however, no stranger than the Swank (1944), Jackson (1946), or the "J" scheme (Pease 1972) for the organization of separate document collections, and retains most of their advantages. For any shelf arrangement decision the first question that must be addressed concerns its purpose. Sayers puts it this way:

> It is essential that the organization on the shelves of a library should reflect the use made of books, not only enabling material on a specific theme to be speedily located, but also subordinating each specific topic to the appropriate general one and showing the most important relationships between subjects (1947, 35).

One does not require sophisticated shelf arrangement simply to locate known works, or even to determine what works are available on a subject. It is clear from Watson and Heim (1984) that SuDoc plays no role in the identification of relevant documents by undergraduates at the University of Illinois. What one hopes to gain is minimization of movement around the collection to find works required, and that at the location of a known useful work there will be other unknown works that may substitute for, or supplement, it. If the work of a corporate author is desired, SuDoc should work well. If material on a topic is desired, SuDoc may well scatter such works. This degree of scatter is currently unknown.

One might presume that the two schemes would collocate material quite differently. It is possible that a collection arranged by LCC might better meet the needs of its users. This could only be the case, however, if materials in SuDoc classes were widely scattered by reassignment into the LCC. Should a rearrangement by LCC result in minimal scatter of SuDoc classes, no particular advantage arises from rearrangement. The presence of scatter would not imply that LCC was more helpful, but its absence would ensure that it could not be. If an hypothesis that no significant scatter exists can be supported, then a decision to move away from SuDoc seems highly questionable. It was hypothesized that no significant scattering of LCC classes occurs over selected SuDoc classes.

## Method

To obtain sample data, four agencies were selected and their records from the Library of Congress MARC File on the DIALOG® system were downloaded. A general sampling of SuDoc numbers was taken by collecting all LC-MARC records that contained SuDoc numbers; arbitrarily chosen sets of 15–20 records were downloaded. This resulted in a broad sample of 528 records, a Library of Congress file of 225 records, an agriculture file of 567, an army file of 364, and a state file of 419.

Records were reduced to integer pairs representing their ranks when sorted in each scheme. These pairs were used to generate rank order statistics and other measures, with the exception of the scatter measures and class counts.

O'Neill, Dillon, and Vizine-Goetz (1987) have introduced measures of distance and scatter for one classification scheme mapped to another. Their scatter measure is the expected number of classes one would search to find an item from the original scheme after conversion, starting at the largest new class and continuing in descending order. Their distance measure is based upon the average change in the number of positions between any two documents after conversion, and resembles standard deviation.

Another possible measure is average minimum document movement (AVMDM). Assuming that a stack is closed and circular, the number of positions a document is displaced by reclassification depends on the direction of the move, except where the new position is the second intercept of a diameter of the circle drawn through the initial position. Consider the documents to be ranked 1 to n by LCC and each document to be represented by an integer indicating its position in a ranking by SuDoc. The AVMDM is the average difference between the new position and the old position of each document on the shelf. When given a difference larger than one-half the number of documents, the minimum document movement is the total number of documents minus that difference; when otherwise, the original difference. For comparability this value can be normalized by dividing by one-half the number of documents in the set under examination.

Such measures show the varying degrees of scatter occuring in separate classes when reorganization takes place, but do not provide a means of comparing the actual degree of change that occurs over the whole scheme. It is unclear what any such value means except relative to other values computed from different classes. However, if the actual number of LC main and subclasses within a SuDoc class is high, relative to the possible number of such classes, that would indicate a generally high scatter. Main classes were determined by counting the number of unique first letters in the LC classification numbers (E and F were considered as one letter). Subclasses were determined by counting the number of unique first two-letter pairs and adding a class for any single first letter. Classes E-F and Z each have only one subclass possible because their notation contains only a single initial letter. There are 21 main classes and 213 subclasses by this procedure. The common rank order statistics, e.g., Kendall or Spearman (Hollander and Wolfe 1973; Kendall 1938; Spearman 1904), also provide a measure of the overall relatedness of two separate orders of the same items.

## Results

The Kendall and Spearman coefficients provide little reason to argue for a relationship between the orders. For the Department of Agriculture, Kendall is .032 and Spearman −.028; for the Department of Army sample the coefficients are .046 and .072; for Library of Congress, −.058 and −.083; for Department of State, .119 and .278. The broad sample Kendall is −.131 and Spearman is −.205. No value is far enough from zero to suggest a serious relationship. In terms of LC classes, in no case were there fewer than 48% of the main classes or 12% of the subclasses present in a SuDoc class.

With the O'Neill, et al. measures, the SuDoc army sample and the broad sample demonstrate relatively high distance and scatter; Library of Congress sample, relatively low distance and scatter. The SuDoc agriculture sample exhibits high distance and low scatter, while the state class has high scatter but low distance.

The normalized values (NAVMDM = AVMDM/.5n) are as follows: agriculture .3744; army .5089; broad .5941; Library of Congress .5150; and state .3853. Thus, a document moves between 37% and 59% of the available positions when converted from SuDoc to LCC.

## Conclusions

These results argue for little relationship between the shelf arrangements generated by SuDoc and LCC. The LC classes in each SuDoc class are numerous and diverse, and the lowest normalized average minimum document movement caused by a switch from SuDoc to LCC is 37% of the available document positions. The hypothesis of no significant scattering by a move from SuDoc is not supported.

As seems intuitively correct, we believe that reorganizing documents to LCC from SuDoc results in significant changes and that a user searching for a topic collocated by the LCC scheme would find the SuDoc arrangement far less than convenient. If usage patterns in a document collection could be categorized as to whether they best fit the subject constraints of LCC or the corporate author constraints of SuDoc, evidence would be available to help determine which scheme would make the most sense for a particular separate collection

## References


Hollander, Myles, and Douglas A. Wolfe. 1973. *Nonparametric Statistical Methods* (New York: John Wiley & Sons), 185–99.

Jackson, Ellen P. 1946. A notation for a public documents classification. *Oklahoma College Bulletin* no. 8:47–53.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrica* 30:81–93.

Larson, Kathleen T. 1979. Establishing a new GPO depository documents in an academic library. *Law Library Review* 72(Summer):490–94.

O'Neill, Edward T., Martin Dillon, and Diane Vizine-Goetz. 1987. Class dispersion between the Library of Congress classification and the Dewey Decimal classification. *Journal of the American Society for Information Science* 38(May):197–205.

Pease, Mina. 1972. The plain J: a documents classification system. *Library Resources and Technical Services* 16(Summer):315–25.

Sayers, W. C. Berwick. 1947. *A Manual of Classification for Librarians*, 4th ed. London: Andre Deutsh.

Spearman, C. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15:72–104.

Swank, Raynard. 1944. A classification for state, county, and municipal documents. *Special Libraries* 35(April):116–20.

Waldo, Michael. 1977. An historical look at the debate over how to organize government documents in depository libraries. *Government Publications Review* 4:319–29.

Watson, Paula D., and Kathleen M. Heim. 1984. Patterns of access and circulation in a depository document collection under full bibliographic control. *Government Publications Review* 1(July–August):269–92.

# Subject Representations in Monographic Records

**Principal Investigator:** Pamela Reekes McKirdy, Assistant Professor, Graduate School of Library and Information Science, Simmons College, Boston

**Abstract**

Title and subject representations in a sample of OCLC monographic records were examined. Term distributions in title and subject fields were counted and the effect of using a stoplist of articles and prepositions was determined. A stemming algorithm was applied to title and subject keywords, and overlap between the resultant term stems was explored. Reduction of the record-to-term explosion rate was calculated. Similar analysis was performed using four subgroups of data: Library of Congress (LC) cataloging, science and technology materials, publications from the 1980s, and records entered since 1983.

Libraries of all types offer access to MARC records through online catalogs. Many catalogs offer indexing by title phrase, subject headings, and keywords in titles, subject headings, and other fields. Keyword indexes are time-consuming to build, require large amounts of storage, and may use significant system resources to provide catalog record access. On the other hand, much has been written about the deficiencies of Library of Congress subject headings (LCSH) for online retrieval, and research into improving subject access has included other approaches such as retrieval through classification or the augmentation of basic bibliographic records. The ability to accommodate cross-references adds to the complexity of determining what access points to implement; the effectiveness of cross-references has not been determined.

A review of the literature indicates that current indexing techniques do not satisfy online catalog users. Ease of use is often traded for improved system capabilities such as restricting searches with commands and logical operators. Adding to existing records may indeed improve access to monographic literature but, for the present, millions of records have not been, and remain unlikely to be, augmented.

Exploration of improved associative indexing techniques and natural language query to online reference database searching may yield results with eventual application to online catalogs as well, but basic exploration of online catalog files must be conducted first to determine whether such extrapolation is warranted. This study examines some of the characteristics of titles and subject representations in monographic records toward that end.

## Methodology

Only records for English-language monographs were used in this study. Juvenile literature, fiction, and partial cataloging records were also excluded. Of an initial 1,239 records randomly selected from the OCLC database, 768 were used in the study. The title field (subfields ǂa and ǂb) for each record was processed word by word. A stoplist of articles and prepositions and other noncontent-bearing words was applied and occurrences of stoplist terms were counted. Nonstoplist terms were processed through a stemming algorithm to remove suffixes. Each of the resultant terms was compared with terms present in fields 650 and 651 (topical and geographical subject added entries) in the record. The following data elements were recorded for each:

- Source record OCLC control number
- Cataloging source
- Year and month of record entered date
- Publication date (008, Date1)
- Broad classification (derived from LC or Dewey)
- Number of subject headings (fields 650 and 651)
- Term subfield source (subfield ǂa or ǂb)
- Term word position within title
- Term keyword position within title
- Code indicating presence or absence of term in subject heading
- Code indicating whether term is an entry term for a subject subdivision
- Code indicating whether term is an entry term for the subject heading
- Tag of the subject heading where term was found
- Indicator value for subject heading where term was found
- Subfield code for subfield where term was found
- Stemmed term

We analyzed both the full term-based data file and a file of data records representing one entry per bibliographic record. Record selection, data record construction, and data counting were all performed using programs written in BASIC and run on an OCLC M310™ Workstation.

## Findings

Single data records for each bibliographic record were reviewed to investigate some of the characteristics of the sample records. About 19% of the records were LC cataloging. Approximately one-third of the records were entered since 1983, one-third between 1979 and 1983, and one-third prior to 1979. Subject heading counts yielded the following distribution: 0, 17.6%; 1, 37%; 2, 27.6%; 3, 12%; 4, 4.3%; 5, .5%; 6, .6%; 7, 0%; 8, .4%.

Use of 250 stoplist terms eliminated 2,753 terms from the 8,188 title terms, a reduction of about 34%. Application of the stoplist to 4,117 subject terms reduced the total by about 6%. Two words, *the* and *of*, were responsible for 40.9% of those terms eliminated by the stoplist. Five more words (*and, in, a, for*, and *to*) added another 35.7%; six others (*on, an, with, by, from*, and *at*) contributed 9% more. Thus, 13 stopwords eliminated 2,350 words. Although use of even these terms may contribute to key-

word searching using Boolean operators and adjacency criteria, significant reduction in storage may be affected through use of a modest stoplist.

Terms used as keywords were further reduced slightly by elimination of terms shorter than three letters (<.5%). The resultant stemmed keywords were examined to determine distribution of term occurrence. High occurrence terms were identified (n > 25): *develop, histor, new, plan, report, state, studi, unit* (in stemmed form).

The occurrence of stemmed title terms in subject headings was determined. Of the 5,399 key terms in titles, 20.6% (1,111) were also present in subject headings. Of those, about half (546, 10.1%) were entry terms for the subject heading. Another 200 (3.7%) were entry terms for a subfield other than the first. Of the subject headings in which title terms were found, 87.4% were in field 650; the remaining 12.6% were in field 651. Almost all (96.4%) of the terms were located in LC subject headings.

To determine the reduction possible due to application of the stoplist and overlap, explosion rates were calculated. The total title term per record factor was 10.66 and the total subject term per record rate was 5.36, or a total of 16 index terms per record. The comparative rates for keyterms were approximately seven title terms and five subject terms per record, or a reduction to 12 (25%) through use of the stoplist. Further reduction due to term overlap yields a rate of 10.6 terms per record (about 12%).

Measures from the four data subsets are compared in table 3.4. Little of the observed variation is significant, although the slightly higher rate for terms found in subject headings in LC cataloging might merit additional exploration, and the higher rate of subject incidence in field 650 for science materials seems quite reasonable.

## Conclusions

Exploration of a fairly small sample of monographic records limits observations to the more obvious trends, but indications are present which appear to be useful. Clearly, mapping of title terms onto subject terms, which results in only 20% overlap, demonstrates that each contributes to access. In the online reference environment, a rate as high as 70% is considered sufficient to conclude that access to both free-text titles and controlled vocabulary subject descriptors is necessary (Henzler 1978). Consideration might be given to combining keyword indexes or providing some guidance to users once the interaction is better understood. Also, further study is needed of the role of cross-references available for subject headings in online catalogs.

Terms found in titles for this project were compared with those available for LC subject headings, but the list only provided the terms alone. It was not possible to identify when cross-references would have served to connect the searcher to the appropriate subject heading. A differently structured study of the terms within the context of the title would prove more useful.

Results from this or similar studies must be combined with observations of the use of indexes provided in online catalogs. Transactional analysis of search sessions is needed to determine how behavior might be influenced or anticipated by better interface design.

## References

Henzler, R. G. 1978. Free or controlled vocabularies: some statistical user-oriented evaluations of biomedical information system. *International Classification* 5:21-26.

**Table 3.4** Summary of Monographic Record Characteristics (%)

| | Term Location | | | | | |
|---|---|---|---|---|---|---|
| Sample | Subject Term | Subject Subdivision Entry | Subject Entry | Field 650 | Field 651 | Title Key Explosion Rate |
| Entire | 20.6 | 13.8 | 10.1 | 87.4 | 12.5 | 7.0 |
| DLC | 27.5 | 17.9 | 13.3 | 88.6 | 11.4 | 5.3 |
| Science | 23.8 | 16.2 | 12.6 | 94.4 | 5.6 | 6.5 |
| New Publications | 23.4 | 15.9 | 11.7 | 88.3 | 11.7 | 7.5 |
| Newly Entered | 21.0 | 13.9 | 9.9 | 88.5 | 11.5 | 7.8 |

# Variations in Personal and Corporate Names in the OCLC Online Union Catalog

**Principal Investigator:** Arlene G. Taylor, Associate Professor, School of Library Service, Columbia University, New York

**Abstract**
This research sampled the OCLC Online Union Catalog to determine the extent to which variations in personal and corporate names are not covered by the Library of Congress authority file (LCAF). Data are analyzed to determine types of variations, dates of entry for both bibliographic and LCAF records, record sources, and rules in force when variants were created.

Research conducted by Elizabeth Fuller (1986) found that 17% of a sample of personal authors taken from the University of Chicago library catalog used more than one form of name in works found in the catalog (not including journal articles, chapters in books, etc.). Each such form has the potential for being used as a heading, as does each variant form used in a journal article or book chapter. We might expect that much of the 17% is taken care of by references. There has been no previous comparable research to show how many corporate names appear in variant forms.

This research is being conducted to determine the extent to which OCLC bibliographic records contain personal and corporate name variants that are not covered by references in the Library of Congress authority file (LCAF).

## Methodology

A sample of bibliographic records was drawn from the OCLC Online Union Catalog. It was not possible to draw a sample of *names* only because OCLC does not have a file that contains only one occurrence of each name appearing in Online Union Catalog records. The sample contains a large representation of prolific personal authors and corporate sponsors. Thus, the findings are examined in terms of the proportions of records that have certain characteristics. This seems valid because libraries usually deal with bibliographic records, not individual names, and need to know what proportions of records may contain one or more names that vary from authorized forms. It also seems valid to deal with many prolific authors and corporate bodies because any given library will deal with more records for those names. It was determined that a sample of approximately 900 items would be needed for a precision of 2% with a confidence level of 95%.

"Authority work" is being completed for every personal name access point (i.e., MARC field 100 or 700) and for every corporate name access point (i.e., MARC fields 110, 111, 710, or 711). Where possible, authority records are found in the online LCAF. Then the Online Union Catalog is searched for records that relate to the same name. Variant forms of each name found both in access points and statements of responsibility are noted, and cataloging methodologies are used to determine which variants represent the same person or body.

For each variant identified, data are being collected in the following categories:

1. Type of variation
2. Appearance of variations in statements of responsibility or notes
3. Number of records that are LC input, member input, or LC created/member input
4. Date of entry of records
5. Existence of an LCAF record and its date of entry
6. Presence of references in the LCAF record (noting especially whether the references are "close matches" with any variants)
7. Rules coding on the LCAF record

## Preliminary Results

Searching to find authority records for the personal and corporate names from the 900 sample records is complete. Forty-one (4.6%) of the records have no personal or corporate name access points. The remaining records yield 900 personal names and 313 corporate names. Authority records were found for 517 personal names (57.4%) and for 257 corporate names (82.1%).

In terms of records, 573 bibliographic records (63.7%) have one or more names for which an authority record has been found in the LCAF. For 498 records (55.3%) an authority record was found for every name on the record. Of the 573 records, 121 have one or more names that differ from the corresponding authority record. This represents 13.4% of all records in the sample or 21.1% of the bibliographic records for which one or more corresponding authority records were found. Records with nonmatching names are (1) created and input by member libraries, 70.2%; (2) created by LC and input by member libraries, 25.6%; (3) created and input by LC, 2.5%; and (4) created and input by the British Library, 1.7%.

There are 17 records in the sample (1.9%) that have names that match an authority record, and which have been "verified" in machine conversion at OCLC with a "≠w cn," but that represent a person who is not the same as the one represented by the authority record. For three of these the correct person is represented on another authority record in a different form. For example, the LCAF contains records for three persons named *Susan Goodman*—one with a birth date, one with a second forename, and one with neither. The sample record lacks the birth date or a second forename but represents a book by the Susan Goodman with a birth date in the authority file. However, the name on the sample record is followed by a "≠w cn," indicating "verification" as being the Susan Goodman with neither birth date nor second forename.

Because the sample was drawn from a database of bibliographic records rather than a file of names, the data for names can be taken only as the literal percentages of the sample. Of the 517 personal names with authority records in the sample, 432 match the corresponding authority record exactly. This represents 83.6% of the personal names with authority records. Eighty-five of these names (16.4% of the personal names with authority records) do not match in form. Of the 257 corporate names with authority records, 214 (83.3% of the corporate names with authority records) match the corresponding authority record exactly. Forty-three of these names (16.7% of the corporate names with authority records) do not match in form.

Analysis is continuing to determine (1) the types of differences involved when nonmatches occur, (2) dates of entry of the nonmatches, (3) the extent to which the nonmatches match a reference, and (4) the extent to which nonmatches are close matches with either authorized forms or references. The next phase of the project is to search the bibliographic database for other records that use the names found in the sample and to determine the extent to which variants of these names exist in the bibliographic file.

### Future Research

A next step for future research would be to study the feasibility of identifying close matches with computer assistance and then using human intervention to determine which close matches are indeed true matches. In addition, comparative studies of variations in other bibliographic databases would be useful. Further analysis of corporate names and initial analysis of geographic names, series, and uniform titles is needed.

### References

Fuller, Elizabeth E. 1986. Variation in personal names in the catalog. Master's thesis, University of Chicago.

# 4

# Distinguished Seminar Series

The Distinguished Seminar Series of lectures, open to interested OCLC staff and members of the library and information science community throughout Ohio and neighboring states, serves as a stimulus for discussion of current research topics. Each year the Office of Research invites six to eight speakers, accomplished in their fields of research, to conduct seminar presentations at the Center in Dublin, Ohio.

Seminar topics range widely, as do the interests of the audience participants. This variety, from statistical theories of retrieval and information impact indicators to communication theory in online retrieval or electronic library applications, represents the many facets of the growing field of information science and ensures the exchange of ideas across traditional academic or professional boundaries.

By sharing new ideas and perspectives and interacting with OCLC staff, the Distinguished Seminar Series speakers play an important role in establishing the research environment at OCLC. Their topics are always fresh, their research current and timely; often their findings are not yet published in the literature. The 1987–88 seminars are reported in the pages that follow.

# Software Architecture to Support Data Translation

Sandra A. Mamrak

There is a pressing need to exchange manuscripts using electronic media, but the wide variety of possible electronic representations makes such exchange difficult," said Sandra Mamrak, reporting on the progress of Project Chameleon sponsored by the Computer and Information Science Department, The Ohio State University.

*"There is a pressing need to exchange manuscripts using electronic media, but the wide variety of possible electronic representations makes such exchange difficult." —Sandra A. Mamrak*

Exchanging machine-readable data requires translation from one form to another. "In many situations, standard forms can be used to reduce the number of translations among different representations," said Dr. Mamrak. The standard form model for translation appears in figure 4.1.

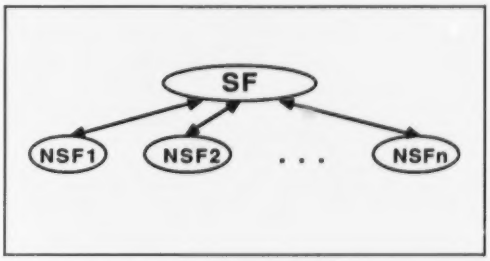



**Figure 4.1** Standard Form Model

Several factors drive the development of translation methodologies: (1) many companies have a significant financial investment in systems that produce nonstandard document representations, (2) there is a large body of data existing in nonstandard forms, and (3) many "standards" are emerging for document representation such as SGML (Standard Generalized Markup Language).

This complicated environment requires development of computer architectures to assist translation and exchange of documents, said Dr. Mamrak. To address this need, Project Chameleon has three goals: (1) to analyze the task of building and using translators, (2) to design and build software architecture to assist all aspects of translation, and (3) to make this architecture accessible to users who are not computer specialists.

### Translation

Translation is essentially the mapping of character strings from one computer or markup language to another. Translation can be "up," i.e., from a nonstandard form (NSF) to a standard form (SF) or "down" (SF to NSF). To support this, the software architecture must take the input and translate it into valid and executable translation code.

"Various problems arise in translation," said Dr. Mamrak, "including loss of important information." If a SF string has no NSF match, this results in a functionality mismatch that requires expert user intervention to resolve. If the SF string matches a NSF string, the translation is uniquely defined. However, a SF string could match more than one NSF string and require user intervention to resolve synonymous strings.

In translation up, the same possibilities exist with one exception. If a NSF string matches more than one SF string, this results in possible ambiguity. For example, commands dictating that text be centered and boldfaced could apply equally to author or title text.

### Software Architecture

"Having defined both the process and the problems of translation up and down, a software architecture can be designed to assist the task," said Dr. Mamrak. "An attribute grammar specifies the translation down, and code is generated for all translation strings."

In the case of a functionality mismatch during translation down, a translator writer provides needed functionality or flags the untranslated string for later expert user intervention. "The translator writer elicits from the user, based on the language of the NSF, a set of mappings to the SF, while protecting the user from the details of the translation architecture," explained Dr. Mamrak. "User input to the translator is automatically transformed into input expected by the system architecture." For synonym resolution, the translation writer chooses one string to complete the down translation.

During translation up, functionality mismatch is handled similarly: either the necessary functionality is provided by the translator writer or the untranslated string is passed to an expert user assisting the process. In the case of ambiguity, the up translator interacts with expert users to resolve the problem. A schematic of the complete architecture, in-

cluding user interfaces, is shown in figure 4.2. At the time of Dr. Mamrak's project report, the translator writer had been implemented, and user interface components were being designed.

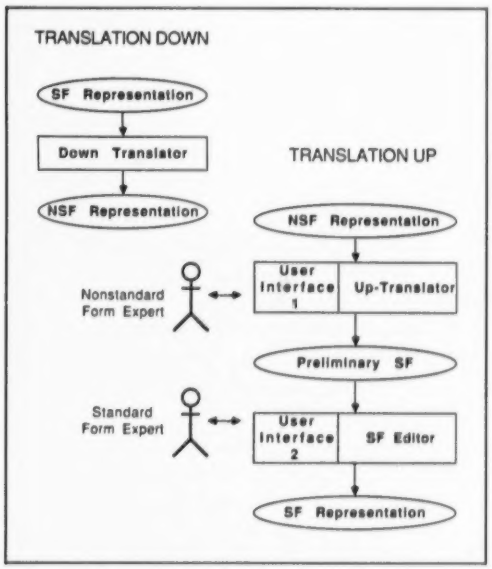



**Figure 4.2** Software Architecture with User Interfaces

Presented July 14, 1987. Dr. Mamrak is Associate Professor, Department of Computer and Information Science, The Ohio State University.

# Changes in the Life Cycle of Information

**Toni Carbo Bearman**

"New measures, or information impact indicators, are needed to assess the effects of changes in the information life cycle on national policy making and the resultant quality of American life," said Toni Carbo Bearman, who presented initial proposals for research that will later be submitted to the National Science Foundation for possible funding.

Dean Bearman's research investigates how changes in the information life cycle affect the use of scientific, technical, and societal information by national governmental policy makers. The goals of her research are to examine changes in the components of the information life cycle, review studies in use of information in policy development, and develop indicators to measure social impact of information on policy formulation. Her research will join data related to scientific, technical, and societal information with national policy development.

## Life Cycle of Information

"The life cycle of scientific, technical, and societal information itself is unchanged," said Dean Bearman. "Information is created, recorded, organized, stored, disseminated, and used. But the components of the life cycle are changing."

The advent of powerful computers, workstations, and extensive electronic data and communications systems has affected how information is processed in all components of the information life cycle from the generation and capture of data to its subsequent transformation, description, storage, access, and use. Advances in miniaturization, rapid developments in software design, new generations of computers, and satellite technology provide vast quantities of raw data that can be examined and manipulated in ways previously impossible. "The greatest supercomputers [now available] are still inadequate for physical, biological, and chemical research," Dean Bearman said. What is now considered a flood of information is likely only a trickle of what the future holds.

Once data are captured and analyzed, scientists can use databases to conduct extensive literature searches, compose research reports electronically, and disseminate information via communication networks. Research, authorship, peer review, and formal publication may be conducted entirely online in an international environment. The effects of these changes on how scientists and engineers access information, conduct research, and disseminate findings are measurable. For example, research indicates that scientists and engineers read more journals than they subscribe to. In 1975, researchers subscribed to an average of 5.4 journals, down from 6.5 in 1977. Conversely, in 1985 researchers read an average of 13 journals, up from 10 in

*"With greater access to more and more varied information, libraries will continue to play an important role in the life cycle of information." —Toni Carbo Bearman*

1977. Moreover, the methods of electronically accessing information have increased. In 1985-86, there were 760 communication networks, up from 608 in 1977, and more than 76,000 libraries of all types participated in one or more networks, up from 32,000 in 1977.

Once published, whether electronically or in printed form, the information is used by two distinct communities, users and intermediaries. Secondary services such as cataloging, abstracting, indexing, translating, and analyzing performed by intermediaries are additional intellectual processes that improve the accuracy, precision, and comprehensiveness of information. Both the original publications and the information produced by intermediaries must then be accessible to users. Increasingly, automated library systems, online databases, and communication networks provide access to the information.

"With greater access to more and more varied information," said Dean Bearman, "libraries will continue to play an important role in the life cycle of information, especially by providing more customized information services to increasingly sophisticated users."

*"The key question is how to make information work better for society. Scientific and technical information are pervasive determinants of modern life." —Toni Carbo Bearman*

### Information Intensive Society

"In an information intensive society, information permeates everything we do in our lives," said Dean Bearman. "Now we are living in a world where we are inundated by information, reliant upon interdependent information and communication services, with less time to cope with technological changes in an increasingly international environment."

To emphasize this point, Dean Bearman reviewed the almost prophetic projections of the 1976 *National Information Policy Report* produced by the National Commission on Libraries and Information Science. The report anticipated: an exponential increase in the volume of information flow; a shrinkage in the time and distance restraints in communication; a greater nationwide dependence on information and communication services; an increase in the interdependence of previously autonomous institutions such as government, education, and industry; new economic, political, and societal values placed on information; a decrease in the "time cushion" between technological changes and their social impact; and greater international exchange within an emerging "global village." These rapid and radical changes made possible through advances in technology affect not only how research is conducted, but also how research findings are used.

Dean Bearman emphasized that "the key question is how to make information work better for society. Scientific and technical information are pervasive determinants of modern life, yet science rarely has an impact on society until it is transferred to some practical application. For example, diseases are not cured by scientific findings, but by the development of drugs coupled with educational campaigns and mechanisms for distributing the drugs, often with substantial funding from government and industry."

### Information Eras

Dean Bearman outlined three distinct eras of information usage. At first, the use of information services was restricted to specific disciplines and oriented toward the advancement of research. The application and interaction of scientific and technical information from multiple disciplines to achieve a clear national mission such as a moon landing introduced the second era. Attempting to solve specific societal problems such as pollution or solid waste disposal moved society into the third information era and more closely tied national science policies with broader public issues.

"Science policy cannot be isolated from public policy because the final policy decisions result in the setting of national priorities and the allocation of resource," she said.

Dean Bearman said that society has entered a fourth era of information usage typified by more individualized and customized information services. Users are able to access information from desk-top workstations and obtain information, either directly or through intermediaries, tailored to specific needs. Once the realm of scientific investigators, online access to information is now available to increasing numbers of policy advisors and legislators.

### Formulating National Policy

"The role of scientific, technical, and societal information in establishing public policy will become increasingly important in the decade ahead," she said.

The process of policy making, much like the life cycle of information, can be identi-

fied. There must be a perception that a problem exists worthy of attention and resources. This is followed by the interaction of interested parties and the formulation of policy recommendations, which must then receive the support of a larger group of policy makers. Legislation formalizes policy and often authorizes the expenditure of funds. The legislation is then implemented and administered, and its consequences are monitored.

"Making decisions about complex issues requires more than the expertise of up-to-date, knowledgeable specialists" who use automated information processing systems, said Dean Bearman. "Most policy decisions at state and national levels are made by generalists, not scientists and engineers." Describing the United States Congress as an information processing organism, she said that "the ability of members of Congress, and Congress as a whole, to effectively process the best information available determines the kind of public policy enacted."

Dean Bearman suggested the following list of possible indicators of changes in the information life cycle that might determine effects on national policy making:

- Number of educational programs for information professionals
- Number of graduates from the above programs and their salaries
- Number of scientists and engineers preparing research papers online
- Number of journals available online
- Number and types of public interest groups providing recommendations and policy briefings to members of Congress
- Number of searches of databases by policy advisors
- Percentage of policy recommendations that are actually proposed for further action
- Percentage of topics recommended by technical experts that become topics of congressional hearings
- Use of new information technology by members of Congress
- Scientific, technical, and societal concerns given high priority by cabinet-level officials
- Number of citizens able to respond to policy makers via electronic networks

Dean Bearman said that these and other information impact indicators will provide accurate, valid, and relevant measures of the complicated relationships among changes in the information life cycle and the impact of those changes on both the policy making process and the resultant quality of American life.

---

Presented September 16, 1987. Dr. Bearman is Dean, School of Library and Information Science, University of Pittsburgh and former Executive Director, National Commission on Libraries and Information Science.

## Tomorrow's Library Today

W. David Penniman

Information transfer is basically a human enterprise; you have to rely on good people as well as good technology" said W. David Penniman, describing the unique blend of library personnel, specialized training, and access technologies that make up the AT&T Information Access Station, a self-contained access point to the AT&T library network.

"The Information Access Station is not a library without walls," said Dr. Penniman, "but a modular library access point that costs less per individual served and delivers more service. Designed to be mobile, flexible, and freestanding, the Information Access Station uses available resources and off-the-shelf technologies to combine the physical, electronic, and human aspects of information delivery."

### Forces Shaping Development

Dr. Penniman cited a host of factors that shaped the development and installation of Information Access Stations throughout AT&T. Chief among them are rising labor and materials costs. Resources for staff and materials, particularly increasingly expensive foreign technical journals, are inadequate to meet rising user demands. Library users are asking for expanded services and access to

*"Information transfer is basically a human enterprise; you have to rely on good people as well as good technology."*
*—W. David Penniman*

more and more kinds of information. In addition, the cost of space for library services is rising and its availability is decreasing. Offsetting these limiting forces are greatly increased computing and worldwide networking capabilities.

Of this was born the Information Access Station. The Access Station is designed for high traffic, easy access areas such as lobbies, atriums, and office cafeterias. It can fit in less than 400 square feet to make use of available office space, too. The station is easy to assemble and move, requires low initial investment and maintenance, and attracts users. Dr. Penniman estimates that each Access Station costs about $10,000 to equip fully.

*"The goals of the Access Station are productivity, quality, timeliness, and value for the library user."*
*—W. David Penniman*

Highly trained librarians with a strong service orientation staff the stations and add to their high visibility. "We select staff who are self-motivated, sales oriented, articulate, good listeners, and experts in reference and library information services," said Dr. Penniman. To ensure adequate service, staff receive intensive training in automated circulation control, online catalog searching, database retrieval techniques, network access procedures, software, hardware, and computer operating systems.

### Access Window

Dr. Penniman calls the Information Access Station "a window to a tremendous amount of service," noting that it is not necessarily what is visible that determines the level of library services. In fact, behind the Access Station lies the entire AT&T library network of 37 libraries and 18 specialized information services. "The Access Station is symbolic of changes in library information services—one-stop shopping," said Dr. Penniman. Users can access the library system from desktop workstations or terminals located in the Access Stations.

Through a computer terminal, Access Station users can connect with local and remote databases to access and obtain information from more than 500,000 AT&T technical documents, more than 500,000 cataloged photographs, the AT&T personnel database, and commercial database providers worldwide. The Access Station also provides current journals, key reference works, access to back copies of journals stored off site, and holdings information for interlibrary loan. The station is staffed by trained experts who assist users with reference questions and online searching. "The goals of the Access Station are productivity, quality, timeliness, and value for the library user," said Dr. Penniman.

By all measurements, the Access Station is meeting these goals. As reported in the *AT&T Library Network 1986 Annual Report*, demand for library services continues to increase but unit costs continue to drop. For example, because photocopying is done in a production mode at a remote facility, the Access Station reduces the cost of photocopying a journal from $35.00 to $5.00. "We want to expand the number of journals held by the library system, not the number of copies," he said, "and we do this by increasing access to materials."

Unit costs also decrease because of cost avoidance, diminished capital replacement costs, increased use of existing materials by more AT&T staff, and improved productivity in the work place.

"We did not want to wait for technology to solve all of the library's problems," said Dr. Penniman. And rather than wait, he introduced "tomorrow's library today." To that end the Information Access Station represents a successful experiment in providing traditional and valued library services in a new and innovative way.

---

Presented November 9, 1987. Dr. Penniman is Director, Library and Information Systems Center, AT&T Bell Laboratories, Murray Hill, N.J.

For a complete description of the AT&T Library Network, see W. David Penniman. 1987. Tomorrow's library today. *Special Libraries* 78(Summer):195-205.

# On the Nature and Function of Explanation in Intelligent Information Retrieval

Nicholas J. Belkin

If the computerized information retrieval systems of the future are to be truly 'intelligent,' they should be modeled on a functional pattern of human-to-human interaction," said Nicholas J. Belkin. Speaking on the nature and function of explanation in intelligent information retrieval, Dr. Belkin described what he called a "joint cognitive system."

"Four components comprise an information system," said Dr. Belkin, "the user, who is goal oriented, the problem that prompts the user to action, the knowledge base, which is generally a collection of texts or citations, and the intermediary, who mediates between the user and the knowledge base to provide an appropriate response to the problem."

### Intelligent Retrieval

The elements of this model system that render it "intelligent" typically reside with the human intermediary, said Dr. Belkin. "The key is providing an *appropriate* response based on an *understanding* of both the problem and the available knowledge bases."

It is this recognized importance of the human intermediary in current information systems that motivates Dr. Belkin's research. "Computerized information systems should provide interaction equivalent to a human intermediary, and explanation is a major function in human information interaction."

Explanation typically found in expert systems is used to justify a conclusion, describe the logic that led to the conclusion, correct user error, or modify user behavior. "This does not cover the range of explanatory activity," said Dr. Belkin. "In human interaction, explanation is used to achieve comprehension, resolve misunderstanding, check agreement, justify actions, persuade, and modify another's state of knowledge."

### Clarity

Dr. Belkin described the user and the information system as partners in a mutual endeavor. "Partners should understand one another, have matching goals and expectations. The user must understand the system and the roles of all the components," said Dr. Belkin. "Attaining this understanding, or 'clarity,' is a general condition of goal-oriented dialog."

In an information retrieval situation, clarity extends to an understanding of what has been done, what is occurring, and what will be done or is intended during the retrieval session. In the sometimes abstract terms of communications research, clarity pertains to the world [the environment in which the information retrieval activity takes place], the task, the system, others, and self. Obtaining clarity is a function of mutual negotiation between the user and the system as each learns about the other. Dr. Belkin said that clarity can take place in the following modes:

- Implicit (generally based on preagreements or postulates)
- Explanation on demand (typically retrospective)
- Explanation during process (often unprompted and offered by the intermediary)
- Process as explanation (awareness of the process is explanation enough)
- Control by others (system limits, often expressed by menu selections)

"However, many factors determine the degree of clarity achieved in an information retrieval session," said Dr. Belkin, who cited (1) the relationship between the parties, (2) familiarity with the system, (3) complexity of the system, (4) complexity of the problem, (5) understanding of the problem, and (6) immediacy or importance of the problem as important variables. "In addition," said Dr. Belkin, "the roles of the parties involved change during the course of information interaction. What is clear is that information systems must be able to adapt dynamically to the user."

*"Computerized information systems should provide interaction equivalent to a human intermediary."*
*— Nicholas J. Belkin*

*"The key is providing an appropriate response based on an understanding of both the problem and the available knowledge bases." —Nicholas J. Belkin*

**Empirical Data**

Dr. Belkins's model for clarity and the modes of explanation is based on audio tapes of actual conversations between users and human intermediaries. The tapes were transcribed and carefully analyzed. Results of these analyses indicate four major goals of clarification: (1) to elaborate the user's expectations of the system, (2) to make a plan of action evident, (3) to make a model of the user and the problem explicit, and (4) to ensure that the retrieval system is appropriate to the user.

Further data analyses revealed four distinctly predominant forms of explanation. Direct explanation, initiated by the intermediary, was generally unprompted and based on an apparent model of the user held by the intermediary. Explanation as description of the process was also frequent; the intermediary told the user what was going on during the retrieval process. Often explanation contained elements of description coupled with an exploration and matching of models (preconceptions or expectations) held by both the intermediary and the user. Whatever its form, explanation was typically unprompted, offered by the intermediary, and assisted the user to participate effectively in the information retrieval process.

Research indicates that the information exchange between user and intermediary is strongly influenced by the preconceptions and expectations of the parties involved. Collecting and analyzing such information allows researchers to develop models of the explanatory process. These models can then be incorporated in intelligent information systems. Models of the clarification process will be used to guide the interaction between user and intermediary and will direct the explanatory process.

Future intelligent information retrieval systems can be designed for direct access by users. While the intermediary function will still be present in the system, it will be conducted by the computer. How, to what extent, by what means, and when explanation is given will be determined by the models of human interaction. The goal will be to imbue information systems with the intelligence now provided by a human intermediary and made apparent through the significant role of explanation. Assuming the characteristics of their human counterparts, these intelligent information retrieval systems will use explanation effectively to achieve the goal, shared with the user, of obtaining appropriate solutions to the information need.

"A real question remains for future research," concluded Dr. Belkin. "How will an interactive user interface, user modeling, and advanced problem description affect information retrieval?"

---

Presented February 16, 1988. Dr. Belkin is Professor, School of Communication, Information, and Library Studies, Rutgers University, New Brunswick, N.J.

## Automatic Sense Disambiguation

"There now is your insular city of the Manhattoes belted round by wharfs as Indian isles by coral reefs commerce surrounds it with her surf."
—*Moby Dick*

Michael Lesk

Rarely does one venture so far into the world of literature when discussing computerized dictionaries and automatic sense disambiguation as did Michael Lesk when he spoke at OCLC.

"The need to disambiguate search terms becomes increasingly important as more and more diversified users search multiple databases with growing frequency," said Dr. Lesk. He highlighted the problems of ambiguous search terms by example: a newswire profile containing the terms *VLSI, chip, semiconductor*, and *integrated circuit* retrieves material containing phrases such as *chocolate chip cookies, chips of plaster*, and *sponsors are chipping in*, clearly not what the searcher intended to retrieve.

Many online databases are not able to distinguish senses in meaning that are patent to the user, and search terms retrieve irrelevant material. Dr. Lesk has investigated this problem and proposes that future electronic information systems, particularly those

providing full-text searching capabilities, replace manual indexing with automatic text processing using computerized dictionaries. "The system should recognize the correct sense of every word as often as possible," he said, "and this could be done using machine-readable dictionaries that are currently available."

Many computerized dictionaries available today, however, provide minimal information. Entries for the word ***house*** in a typical electronic dictionary, for example, contain a few lines at best, about 30 bytes of information. This compares with the same entry in *Webster's Seventh New Collegiate Dictionary* which provides about 1,200 bytes of data. The same headword in the *Oxford English Dictionary* extends for 8.5 columns for the noun (47,000 bytes) and 1.25 columns for the verb (7,000 bytes). Compounds of the word ***house*** extend for 14 columns (78,000 bytes).

### Sense Overlap

Dr. Lesk asserts that sheer bulk of dictionary information and a process of detecting sense overlap reduces ambiguous terms. The theory of sense overlap uses machine-readable dictionaries to compare meanings of terms or phrases such as *pine cone*. In the *Oxford Advanced Learner's Dictionary of Current English* the word *pine* has two major senses: (1) a kind of evergreen tree with needle-shaped leaves, and (2) waste away through sorrow or illness; *cone* has three distinct definitions: (1) a solid body which narrows to a point, (2) something of this shape whether solid or hollow, and (3) fruit of certain evergreen trees. The words *evergreen* and *tree* are common to two of the definitions. "A program could guess that when the phrase *pine cone* appears the correct meaning is the fruit of the tree," he said.

The method is not without limitations, however, admitted Dr. Lesk. Phrases such as *time flies like an arrow* and *fruit flies like a banana* pose difficulties in natural language processing. In this example, Dr. Lesk's program was able to correctly distinguish the different meanings of the word *flies* but chose the wrong sense of *arrow*.

The advantages of this technique, said Dr. Lesk, are that it is nonsyntactic and not reliant upon global information. "Sense overlap is a useful complement to syntactically based sense resolution. For example, a syntactically based system could distinguish the meaning of the word ***hawk*** in the phrase ***I know a hawk from a handsaw*** because the verb form (to peddle goods by calling out) cannot appear immediately after an article." Dr. Lesk points out, however, that often sense overlap can resolve ambiguity when syntax alone cannot.

In addition, the sense overlap method does not rely on global information. In the passage from *Moby Dick*, the program correctly distinguishes the meaning of the word *reefs* in the phrase *coral reefs*. The more frequent meaning of *reef* in Melville's classic is *to reduce the size of the sail*. If the sense overlap program relied solely on frequency counts, it would select the incorrect meaning.

### Variables of Success

"The more information a machine-readable dictionary provides, the greater the success in discovering sense overlap," said Dr. Lesk. "I think the length of entry is the dominating factor in determining the correct sense." The relative sizes of several dictionaries appear in table 4.1. Other factors to be considered include compound words, idiomatic usage, the inclusion of examples, and the range of words examined. Dr. Lesk encourages more research and experimentation in these areas.

*"The system should recognize the correct sense of every word as often as possible."*
*—Michael Lesk*

**Table 4.1** Dictionary Size

| | OALDC | W7 | CED | OED |
|---|---|---|---|---|
| Megabytes | 6.6 | 15.6 | 21.3 | 350 |
| Headwords | 21,000 | 69,000 | 85,000 | 304,000 |
| Senses | 36,000 | 140,000 | 159,000 | 587,000 |
| Bytes/ Headword | 290 | 226 | 251 | 1,200 |

OALDC = *Oxford Advanced Learner's Dictionary of Current English*
W7 = *Webster's Seventh New Collegiate Dictionary*
CED = *Collins English Dictionary*
OED = *Oxford English Dictionary* (estimated)

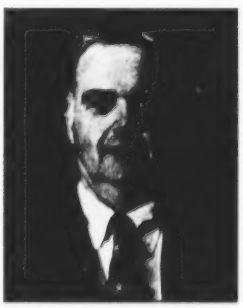
Returning again to the passage from *Moby Dick*, Dr. Lesk raised both the problem and the promise of literary allusion and figurative language. "When Melville wrote 'your insular city of the Manhattoes' did he mean only 'surrounded by water' or also 'narrow minded'? " wondered Dr. Lesk, concluding that both meanings are present. Dr. Lesk envisions the possibility of automatically examining literature using machine-readable dictionaries with sense overlap parameters

defined by a particular literary age to discover figurative usage. "The program could find text having two different, yet internally consistent, meanings. This is perhaps the most exciting future possibility of this work."

Presented March 8, 1988. Dr. Lesk is Division Manager of Computer Sciences Research, Bell Communications Research, Morristown, N.J.

# Probability Theory in Information Retrieval

*"The system can improve the ordering of documents presented to the user based on the probability of user satisfaction."*
*—William S. Cooper*



William S. Cooper

Probability theory has long been a paradigm for research in information retrieval, but now it is time to start implementing this design philosophy in actual systems," said William S. Cooper.

"Because it is probabilistic whether a particular document will satisfy a user's request, the essential action of a document retrieval system in response to a search request is to rank the documents for the user in order of decreasing estimated probability of relevance to the user's information need," said Dr. Cooper, describing what he called the "probability ranking principle." Dr. Cooper advocates using statistical theory to improve information retrieval (IR) systems.

## Retrieval Scenarios

Dr. Cooper explained probability theory by describing three hypothetical IR systems. Probability ranking may be performed by computer, human indexers, or both. "Objective probability estimates are assigned by computer," said Dr. Cooper, "and subjective estimates are assigned by humans."

**Objective probability ranking.** In a system with objective probability ranking, each document is indexed with unweighted descriptors such as subject terms. The system responds to a search request by retrieving, *in arbitrary order*, all documents bearing the search term. No probability ranking occurs and the user has no clue as to which documents are more likely to be of interest.

Probability ranking can be introduced to a system if the user can examine each document retrieved and indicate to the system whether it is relevant to the information need. "Over time, based on accumulated user responses, the system can improve the ordering of documents presented to the user based on the probability of user satisfaction," said Dr. Cooper. "For each document, this probability is based on the proportion of past users submitting the same search term who were satisfied with the document."

**Subjective probability ranking.** For systems based on subjective probability estimation, human indexers assign relative values or weights to terms associated with a document when it is indexed. These probability estimates are really just guesses of the proportion of users who may be satisfied with a particular document in response to a given search term. These estimated probabilities function in place of actual user feedback which, in this scenario, is not obtained by the IR system.

"If the indexers' guesses are better than random, then the ordering of documents is better than arbitrary and the likelihood of relevancy is improved," said Dr. Cooper. "But 'only better than random' is a weak requirement for accuracy."

**Combined probability ranking.** The third scenario combines objective and subjective ranking. Indexers assign subjective probability estimates when documents are indexed, but user feedback enables the computer automatically to recalculate probability estimates based on users' experiences over time.

IR systems using probability estimates would be sensitive to shifts in the demographics of the user population. Relevancy values obtained from a highly specialized user group would skew such a system. Documents ranked by these probability estimates may prove irrelevant to a general user. "A natural refinement would be to assign weights based on demographics," said Dr. Cooper. Probability estimates could be calculated based on information provided by the user or stored in a user profile.

### Utility vs. Relevancy

"We really ought to be talking about the utility of a document," said Dr. Cooper, introducing yet another variable into the IR scenario. "There is often a conflict between relevancy and utility. The probable relevance of a document retrieved by a search request may be high, but if the user has already read the document, its utility to the user may be low.

"Relevance, then, should be defined as the relationship between a user's information need and any particular document," he said. Post-coordinated Boolean search operations can be used to eliminate unwanted documents. But Dr. Cooper recommends "intra-search" adjustment of probability rankings whereby user feedback is used by the system during a single search session to improve user satisfaction. "To this end," said Dr. Cooper, "it's time to move probability theory into practice."

Presented April 8, 1988. Dr. Cooper is Professor, Library and Information Studies, University of California, Berkeley.

## Information Retrieval by Plausible Inference

W. Bruce Croft

Despite twenty years of advances, new approaches are required to improve substantially the performance of information retrieval systems," said W. Bruce Croft. "Much current research attempts to make small improvements in retrieval precision," said Dr. Croft, "without considering new possibilities for information retrieval."

Dr. Croft said the major function of an information retrieval (IR) system is to locate documents of interest to the user. But the user often gets the *system's idea* of what documents ought to be of interest to the user, and there is often a low degree of correspondence between the user's information need and the set of retrieved documents.

The user approaches a document set with an information need or problem which must be represented by a query. Similarly, the contents of documents are usually represented by index terms that are manually assigned by human indexers. Typically, query terms are compared with index terms to determine a set of documents for retrieval. "When the representations are formulated, both the queries and the index terms, there is a high possibility for error and imprecision," said Dr. Croft, "and the likelihood of satisfying the user's information need is diminished."

### Recent Developments

Recent IR research has focused on statistically based search strategies such as vector space, probabilistic, and cluster-based models. Research has shown that similar performance among these approaches produces a low degree of relevant document overlap, meaning that many potentially relevant documents are not retrieved by even the more sophisticated systems.

Moreover, different document representations based on manual indexing, automatic indexing, and citation methods retrieve different sets of relevant documents. According to Dr. Croft, research has shown that the performance of a particular type of representation varies depending on the type of query, and the ability to improve performance of current methods is limited. What is clear is that "if you have a simple document representation you can't build complex search strategies," said Dr. Croft.

### The Next Step

"What we need to improve retrieval performance are some new approaches to IR," said Dr. Croft, "such as multiple strategies and representations for each query." He proposes combining the features of probabilistic, cluster-based, and citation-based representation with browsing capabilities to achieve an overall assessment of the relevance of a document. "But you can't take the user out of the system," cautioned Dr. Croft. "You're expecting too much if you think a system can perform as well as a user to determine the relevance of a document," he said.

Dr. Croft went on to describe the I$^3$R project at the Computer and Information Sci-

*"What we need to improve retrieval performance are some new approaches to IR." —W. Bruce Croft*

ence Department, University of Massachusetts, Amherst. The I³R system simultaneously combines various methods of assessing probable document relevancy using inference.

The system is built on the notion of constrained spreading activation. Search requests begin at different points in the document set and then spread. The search begins with documents ranked highest using probability schemes and then spreads using nearest neighbor and citation links. Plausibility constrains the spreading searches using probabilistic logic and semantic nets. When the searches collide on a single document, the weight of the search is reinforced. Activation of a document is a function of the activation of the starting document, the type of link between documents, and the previous activation of the document by another search, if any. Documents activated by many searches are retrieved. The relationship of terms, concepts, and documents in the I³R database is shown in figure 4.3, and the concept of spreading activation is shown in figure 4.4.

*"You're expecting too much if you think a system can perform as well as a user to determine the relevance of a document."*
*—W. Bruce Croft*

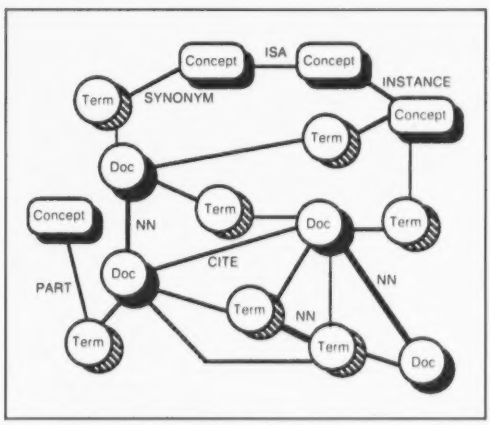



**Figure 4.3** I³R Knowledge Base

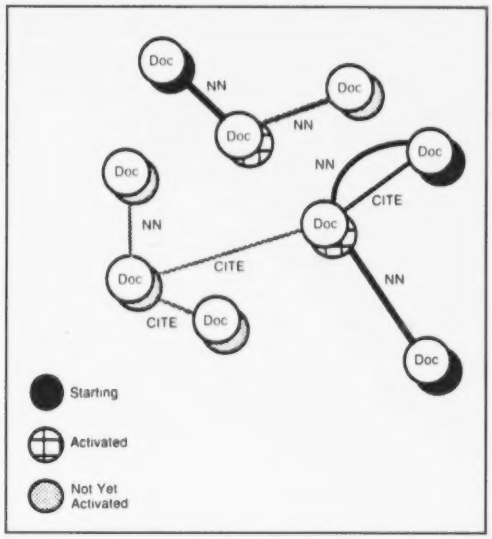



**Figure 4.4** Spreading Activation

Using these techniques, Dr. Croft achieved a 15% increase in precision relative to the best current statistical methods. Yet Dr. Croft believes that adding other procedures will further increase retrieval effectiveness. Natural language processing, interactive user interfaces, automatically generated statistical queries, frame-based representations, complex document objects (e.g., figures, tables, images), and browsing all add additional evidence for retrieval based on plausible inference.

Dr. Croft seeks to find the optimal combination of inferential evidence to assist retrieval, but wonders to what extent evidence from multiple sources can be combined validly. "Are we mixing apples and oranges?" he asked. Despite the complexity of numerous inferential sources for determining document relevancy, Dr. Croft's vision remains clear: "There is no such thing as knowledge representation for the sake of it—it's for IR."

Presented May 3, 1988. Dr. Croft is Associate Professor, Department of Computer and Information Science, University of Massachusetts.

# Programs

## Visiting Scholar:
### A Sabbatical Research Position

The OCLC Visiting Scholar post brings experienced scientists, educators, and administrators with demonstrated research capabilities to OCLC to conduct research with OCLC staff, facilities, and data resources. The Visiting Scholars conduct research in areas of mutual interest to themselves and OCLC so that both benefit from a close working relationship.

#### Previous Visiting Scholars

*1987/88* **Elaine Svenonius,** Professor, Graduate School of Library and Information Science, University of California, Los Angeles

*1987/88* **Paul B. Kantor,** President, Tantalus, Inc., Cleveland, Ohio

*1986/87* **Francis L. Miksa,** Professor, Library and Information Science, University of Texas at Austin

*1985/86* **Martin Dillon,** Professor, Library Science, University of North Carolina, Chapel Hill

*1985* **Terry Noreault,** Assistant Professor, Information Science, University of Pittsburgh

*1984* **Allan Pratt,** Professor, Graduate Library School, College of Education, University of Arizona

*1983/84* **Mary Jo Lynch,** Director, Office of Research, American Library Association, Chicago, Ill.

*1982/83* **Nancy B. Olson,** Associate Professor, Audiovisual Cataloger, Mankato State University, Mankato, Minn.

*1981/82* **Donald J. Sager,** Administrative Libraries, Elmhurst Public Library, Elmhurst, Ill.

*1980/81* **Pauline A. Cochrane,** Professor, School of Information Studies, Syracuse University

*1979/80* **William E. McGrath,** Associate Professor, School of Information and Library Studies, State University of New York at Buffalo

*1978/79* **Edward T. O'Neill,** Professor of Library Science, Case Western Reserve University, Cleveland, Ohio

#### Qualifications

The Visiting Scholar must have a Ph.D. degree or equivalent training and substantial experience in directing and conducting research in one or more fields of interest to OCLC such as library, information, or computer science; applied mathematics; statistics; psychology; and human factors. The position is not restricted to citizens of the United States.

#### Nature of Appointment

The appointee is expected to conduct research that focuses on problems of significance to the library and information science community; the research need not be specific to OCLC development and production activities. OCLC expects that research results will be published in the open literature. All publications stemming from the research conducted while at OCLC are attributed to the Visiting Scholar; acknowledgment of OCLC support is required, and coauthorship is expected when OCLC staff make significant contributions to the research effort.

#### Terms of Appointment

The Visiting Scholar is usually appointed for one year. Part or all of the appointee's salary and benefits may be paid by OCLC during the term of appointment.

#### Research Assistants

The Visiting Scholar may be assigned one or two research assistants with whom to collaborate in conducting the proposed research. The period of appointment for the research assistants will run concurrently with that of the Visiting Scholar. Research assistants must be enrolled in an accredited graduate program in an area related to the Visiting Scholar's research interests. OCLC expects that the research conducted by the research assistants during the appointment period will result in completion of their dissertations or publication in the open literature. Stipends for the appointment period will be available; travel expenses to enable research assistants to attend relevant conferences and workshops may also be provided. Selection of the research assistants is made by the Visiting Scholar in consultation with the Director of the Office of Research.

## Postdoctoral Fellowship

The OCLC Postdoctoral Fellowship provides the opportunity for researchers at the start of their careers to conduct research with OCLC staff, facilities, and data resources. The Postdoctoral Fellow will normally conduct research in areas of mutual interest to OCLC and the researcher.

#### Qualifications

The Postdoctoral Fellow must have a Ph.D. degree or equivalent training in one or more fields of interest to OCLC such as library, information, or computer science; applied mathematics; statistics; psychology; and human fac-

tors. The position is not restricted to citizens of the United States.

### Nature of Appointment

The Postdoctoral Fellow is expected to conduct research that focuses on problems of significance to the library and information science community. Generally, the Postdoctoral Fellow conducts research in one of OCLC's ongoing projects under the direction of a research scientist.

All publications stemming from the research conducted while at OCLC will be attributed to the Postdoctoral Fellow. Acknowledgment of OCLC support is required, and coauthorship is expected when OCLC staff make significant contributions to the research effort.

### Terms of Appointment

The Postdoctoral Fellow is normally appointed for one year. OCLC offers a competitive salary and benefits, and relocation assistance is available.

### Previous Postdoctoral Fellows

*1987/88* **Jeanette Drone,** Ph.D., Library and Information Science, University of Illinois at Urbana-Champaign, 1984

*1986/87* **Jeanette Drone,** Ph.D., Library and Information Science, University of Illinois at Urbana-Champaign, 1984

*1985/86* **Chandra Prabha,** Ph.D., Library and Information Science, University of Illinois at Urbana-Champaign, 1984

*1985* **Padmini Das-Gupta,** Ph.D., Information Transfer, Syracuse University, 1985

*1983/84* **Diane Vizine-Goetz,** Ph.D., Library and Information Science, Case Western Reserve University, 1983

## Research Assistants

The OCLC research assistant position enables full-time graduate students to participate in research at OCLC in Dublin, Ohio, while pursuing an academic degree. Program participants may pursue independent research using OCLC facilities during nonwork hours to complete academic requirements.

### Research Assistant Activities

Research assistants play an important and significant role in research activities and work closely with research scientists to

- Assist in the design, implementation, and execution of research projects
- Conduct literature searches
- Collect and analyze data
- Analyze and write computer programs
- Assist the preparation of research proposals, reports, and publications

### Qualifications

Candidates must be enrolled full time in a graduate degree program at an accredited university. Degree programs may include, but are not limited to, the following general subject areas:

- Computer science
- Information science
- Library science
- Statistics
- Psychology
- Linguistics
- Communications
- Industrial and systems engineering

### Terms of Appointment

OCLC provides research assistants a monthly stipend and pays tuition fees for each academic term the position is held. The stipend and tuition payments are subject to applicable federal, state, and local taxes, and fringe benefits are not included. Generally, research assistants work 20 hours per week, but working hours are flexible.

### Reappointment

Renewal of the position depends upon both academic and OCLC performance. Reappointment to succeeding academic terms is considered prior to the end of a current term.

# Additional Program Information

## Environment, Travel, and Applications Information

### Environment

Resources available to the Visiting Scholar, the Postdoctoral Fellow, and research assistants include access to the OCLC Online Union Catalog and other online information providers as well as other data related to the OCLC database and communications network. Hardware such as XEROX® Sigma 9, Apollo®, Tandem™, SUN, VAX™, MicroVAX™, Pyramid Technology™, IBM and Apple computers and software (including utility routines) apart from the OCLC operations computer system are available for research purposes. OCLC staff are available for consultation. Secretarial, editorial, and graphics support are also provided.

### Travel

Relevant travel and business expenses connected with the Visiting Scholar or Postdoctoral Fellowship appointment and the research assistant position are reimbursed by OCLC, when approved by the director of the Office of Research.

### Applications

Candidates for the Visiting Scholar or Postdoctoral Fellowship appointment should submit a letter of interest, including a current curriculum vitae, date of availability, and specific research interests to:

**Director, Office of Research**
**OCLC Online Computer Library Center, Inc.**
**6565 Frantz Road**
**Dublin, Ohio 43017-0702**

The application deadline is February 1. Candidates are selected by April 1 for appointment in either July or September.

Candidates for research assistantships should submit a letter of interest, a current curriculum vitae or resume, and the date of availability to:

**Human Resources Department**
**OCLC Online Computer Library Center, Inc.**
**6565 Frantz Road**
**Dublin, Ohio 43017-0702**

Applications are accepted at any time. OCLC is an Equal Opportunity Employer.

# OCLC Staff

## Biographies

**John R. Bonneau,** Programmer/Analyst, Office of Research

B.S., Computer Science and Human Factors, Kearney State College, Kearney, Neb.

Mr. Bonneau assists the design and implementation of the user interface for the Advanced Interface Management project.

**John A. Bunge,** Associate Research Scientist, Office of Research

B.A., Music, Macalester College, St. Paul, Minn.; M.A., French Literature, Miami University, Oxford, Ohio; M.S., Mathematical Statistics, The Ohio State University; pursuing a Ph.D. degree in mathematical statistics at The Ohio State University

Mr. Bunge is project manager for the Selecting Bibliographic Records project and statistical consultant to the Database Quality project. He is interested in identifying significant bibliographic variables and in mathematical theory of classification.

**Cheryl Chirakos,** Research Assistant, Office of Research

B.A., English Literature, Evangel College, Springfield, Mo.; M.L.S., Kent State University, Kent, Ohio

Ms. Chirakos maintains a bibliographic database and publication control system.

**Mei-chin Chung,** Programmer, Office of Research

B.S., Zoology, National Taiwan University

Ms. Chung provides programming support for the GraphText project. She has experience developing textual databases and interactive retrieval systems.

**Jack Cox,** Research Assistant, Office of Research

B.S., Computer Information Science, The Ohio State University

Mr. Cox is a programmer for the Library of Congress Class M Subset project. His interests include system software and operating systems.

**J. Charles Crabb,** Research Assistant, Office of Research

B.S., Computer and Electrical Engineering, Purdue University; pursuing a master's degree in computer and information science at The Ohio State University

Mr. Crabb is designing and implementing an interface tool kit based on object-oriented and event-driven paradigms.

**Mark Crook,** Consulting Systems Analyst, Office of Research

B.A., English, The Ohio State University; pursuing a master's degree in library science through the Kent State University/Columbus (Ohio) extension program

Mr. Crook is a member of the Collection Analysis project and manager of the Office of Research Information Services. He is interested in fourth generation languages (4GL), prototyping, database systems, and statistics.

**Martin Dillon,** Director, Office of Research

B.S., English, Canisius College, Buffalo, N.Y.; Ph.D., English, State University of New York at Buffalo

Dr. Dillon directs the Office of Research and is project comanager for the Advanced Interface Management project and is a member of the Collection Analysis system project team. Previously, he was a Professor of Library Science at The University of North Carolina, Chapel Hill. He designed and directed the development of numerous automated systems, particularly in the areas of library applications and information retrieval.

**Jeanette M. Drone,** Postdoctoral Fellow

B.S., Music Education, George Peabody College for Teachers, Nashville, Tenn.; M.Mus., University of Michigan; M.L.S., George Peabody College for Teachers; Ph.D., Library and Information Science, University of Illinois, Urbana-Champaign

Dr. Drone is manager for the Library of Congress Class M Subset project. She is interested in the measurement and evaluation of library collections and services, cataloging, music bibliography, and the development of hypermedia reference systems. Previously, she was Visiting Assistant Professor at the University of Illinois Graduate School of Library and Information Science and, from 1969 to 1980, Music Librarian at Memphis State University. Beginning Fall 1988, Dr. Drone is visiting Assistant Professor, Graduate School of Library and Information Science, University of California, Los Angeles.

**Tony Ershadi,** Programmer/Analyst, Office of Research

B.S., Computer Science, The Ohio State University

Mr. Ershadi assists research funded jointly by the Council on Library Resources, The University of Michigan, and OCLC to increase the accessibility of Library of Congress subject headings in online bibliographic systems.

**Jeffrey L. Gerckens,** Research Assistant, Office of Research

B.S., Computer and Information Science; M.S., Computer Science, The Ohio State University

Mr. Gerckens is developing software for the Graph-Text project.

**Nancy M. Gould,** Research Assistant, Office of Research

B.S., Chemical Engineering, The University of Michigan; pursuing a master's degree in statistics at The Ohio State University

Ms. Gould is statistical consultant to the Office of Research Information Services project. Previously, Ms. Gould worked for General Motors Corporation as a development engineer.

**William F. Guthrie,** Research Assistant, Office of Research

B.A., Mathematics, Case Western Reserve University, Cleveland, Ohio

Mr. Guthrie provides analytical and statistical support. He is interested in robust statistical estimation.

**Sehchang Hah,** Research Assistant, Office of Research

B.A., Psychology, Seoul National University; M.A., Human Performance Psychology, The Ohio State University; pursuing a Ph.D. degree in human performance psychology at The Ohio State University

Mr. Hah has worked on analyses of online search behavior and interlibrary loan, an experimental study on enhancing CRT display, and the Graph-Text project. Previously, he was a research associate in human performance psychology at The Ohio State University.

**John C. Handley,** Associate Research Scientist, Office of Research

B.S., M.S., Mathematics, The Ohio State University

Mr. Handley is manager for the Document Facsimile Transmission Studies project, member of the Duplicate Detection project, and contributor to the Automated Document Structure Analysis project.

**Craig Henderson,** Research Assistant, Office of Research

B.S., Computer Information Science, The Ohio State University; pursuing a master's degree in computer information science at The Ohio State University

Mr. Henderson is developing programs for the Automated Document Structure Analysis and Connectionist Models projects. He also helped develop the HyperBach hypermedia reference system. Previously, he was a consultant for Riverside Methodist Hospital, Columbus, Ohio.

**Thomas B. Hickey,** Consulting Research Scientist, Office of Research

B.S., Physics, State University of New York at Stony Brook; M.L.S., State University College of Arts & Sciences, Geneseo, New York; Ph.D., Library Science, University of Illinois, Urbana-Champaign

Dr. Hickey is manager for the Graph-Text and Information Display projects. His interests include digital typeface generation and display, typesetting language translation, the display of documents with tables and graphics, and interactive bibliographic retrieval systems. He has conducted research on the detection of duplicate bibliographic records, the use and design of search keys in the OCLC Online System, matching personal author names, the use of SGML for interactive document representation, and interpreting page description languages.

**Charles T. Huff,** Programmer/Analyst, Office of Research

M.S., Computer Science, Indiana University

Mr. Huff provides programming support for the Automated Document Structure Analysis project. He is interested in natural language interfaces. Previously, he was instructor of computer and information science, The Ohio State University.

**Paul B. Kantor,** Visiting Scholar

Ph.D., Physics, Princeton University

Dr. Kantor investigated the Scholars Cross-Reference System. He is president of Tantalus, Inc., and conducts research and provides consultation on the economics of information ranging from library management to design of detection and retrieval systems.

**Martha Gordon Lindeman,** Research Scientist, Office of Research

Ph.D., Psychology, Harvard University

Dr. Lindeman co-manages the Advanced Interface Management project. Prior to joining OCLC in 1987, she was Principal Investigator/Project Manager for an Air Force subcontract on a project to define the functional specifications for an intelligent information retrieval system. Her research interests focus on how people process information, particularly as it relates to computer interface design. She is a member of the Human-Computer Interaction Standards Committee of the Human Factors Society, and is also the First President of the Central Ohio Chapter of the Society.

**Susan McCabe,** Research Assistant, Office of Research

B.A., M.A., Anthropology, The Ohio State University; pursuing a Ph.D. degree in anthropology at The Ohio State University

Ms. McCabe gathers research and bibliographic information and assists the Library School Research Equipment Support and Library and Information Science Research Grant programs.

**Michael J. McGill,** Vice President, Research, Technology Assessment, and Development

B.A., Communications, Michigan State University; M.A., Communications Research, Syracuse University; Ph.D., Computer and Information Science, Syracuse University

Dr. McGill is responsible for technology assessment, standards activities, product development, and the conduct and integration of research. Previously, he held positions as Program Director for Information Science at the National Science Foundation; Senior Computer and Information Science Advisor for the U.S. Environmental Protection Agency; and Assistant and Associate Professor and Assistant Dean for Research in the School of Information Studies at Syracuse University.

**Bruce L. Mortland,** Programmer/Analyst, Office of Research

B.S., Computer Technology, Youngstown State University, Youngstown, Ohio

Mr. Mortland is studying new technologies, artificial intelligence, and human-computer interaction with specific application to the Advanced Interface Management project. Previously, he was employed by several systems consultants in Youngstown, Ohio.

**Lora J. Nelson,** Senior Programmer/Analyst, Office of Research

B.S., Computer and Information Science, The Ohio State University

Ms. Nelson is developing typesetting software for the Graph-Text project. Previously, she was a programmer for the Political Science Department at The Ohio State University.

**Edward T. O'Neill,** Senior Research Scientist, Office of Research

B.A., Liberal Arts, Albion College, Albion, Mich.; B.S.I.E (Industrial Engineering), M.S.I.E., Ph.D., Purdue University

Dr. O'Neill is project manager for the Database Quality project and related subprojects, including authorities and authority control, the investigation of error correction techniques, collection analysis, index evaluation, and database organization. Previously, he was Dean of the Matthew A. Baxter School of Information and Library Science at Case Western Reserve University. Prior to that, he was an Associate Professor in the School of Information and Library Studies at the State University of New York at Buffalo.

**Ildiko E. Pallos,** Research Assistant, Office of Research

M.A., Experimental Psychology, The Ohio State University; pursuing a master's degree in library science through the Kent State University/Columbus (Ohio) extension program

Ms. Pallos maintains an in-house bibliographic database for the Office of Research.

**Debra A. Pawlicki,** Research Assistant, Office of Research

B.A., History, Illinois Benedictine College, Lisle, Ill.; M.B.A., DePaul University, Chicago

Ms. Pawlicki is currently writing design documentation and testing programs using the D language developed for the Advanced Interface Management project. Previously, she was a course developer and instructor for AT&T Technologies and a communications systems representative for AT&T Information Systems.

**Kym Pocius,** Research Assistant, Office of Research

B.S., Psychology, The Ohio State University; M.S., Psychology, Virginia Polytechnic Institute and State University, Blacksburg, Va.; pursuing a Ph.D. degree in psychology at The Ohio State University

Ms. Pocius is a programmer and assistant for the Advanced Interface Management project. Her interests include brain topography and its relationship to cognitive styles.

**Chandra G. Prabha,** Research Scientist, Office of Research

B.S., Education, M.L.S., University of Wisconsin, Madison; Ph.D., Library and Information Science, University of Illinois, Urbana-Champaign

Dr. Prabha is manager for the Managing Large Retrievals project. She has also conducted nonfiction book use studies in public and academic libraries. Previously, she conducted research in bibliometrics. Her library experience includes appointments at the University of Wisconsin, Madison; Milwaukee Public Library; International Crops Research Institute for Semiarid Tropics; University of Hyderabad, India; and the American Studies Research Centre.

**Michael J. Prasse,** Research Assistant, Office of Research

B.A., Mathematics and Psychology; M.A., Experimental Psychology, University of Southern California; Ph.D., Cognitive Psychology, The Ohio State University

Mr. Prasse is a member of the Advanced Interface Management project and conducts research on human factors issues pertaining to the presentation of information on a display screen. He also evaluates the cognitive and perceptual implications of reading text from a terminal. Previously, he was a teaching assistant in the Psychology Department and an administrative assistant for the Graduate School at The Ohio State University.

**Mahendra A. Ramachandran,** Research Assistant, Office of Research

B.S., Computer Science, The Ohio State University; pursuing a master's degree in computer science at The Ohio State University

Ms. Ramachandran is a member of the Advanced Interface Management project and is modifying the D programming language.

**Karen J. Rankey,** Senior Programmer/Analyst, Office of Research

B.S., Computer and Information Science, The Ohio State University

Ms. Rankey analyzes the Online Union Catalog for the Database Quality project. Previously, she was a programmer for The National Center for Research in Vocational Education, Columbus, Ohio.

**Joseph W. Reilman,** Research Assistant, Office of Research

Pursuing a bachelor's degree in accounting and computer information systems, The Ohio State University

Mr. Reilman provides programming support in the C language.

**Anthony M. Repka,** Senior Programmer/Analyst, Office of Research

B.A., Geography; M.S., Computer Science, Bowling Green State University, Bowling Green, Ohio

Mr. Repka is designing and implementing features of the D language for the Advanced Interface Management project. He is also system administrator for the Sun-3 computer network. Previously, he was a programmer for the Department of Psychology at Bowling Green State University.

**Duane A. Rice,** Research Assistant, Office of Research

B.S., Industrial Design, The Ohio State University; pursuing a master's degree in industrial design at The Ohio State University

Mr. Rice assisted development of the HyperBach hypermedia reference system. Currently, he is developing a prototype system to display equivalent bibliographic records and is consultant to the Advanced Interface Management project on human-computer interaction.

**David J. Stephens,** Senior Programmer/Analyst, Office of Research

B.S., Industrial Engineering, The Ohio State University; pursuing a master's degree in business administration at The Ohio State University

Mr. Stephens provides support for systems programming, hardware systems integration, and applications development.

**Taylor Surface,** Programmer/Analyst, Office of Research

B.S., Computer Science, The Ohio State University

Mr. Surface is a programmer for a fully automated, syntactically based text indexing system.

**Elaine Svenonius,** Visiting Scholar, Office of Research

Ph.D., Library Science, University of Chicago

Dr. Svenonius is manager of the Clustering Equivalent Bibliographic Records project. She is Professor, Graduate School of Library and Information Science, University of California, Los Angeles.

**Mary C. Taylor,** Research Information Specialist, Office of Research

B.A., History, University of Illinois; M.L.S., Kent State University, Kent, Ohio

Ms. Taylor is program manager for the Library and Information Science Research Grant program. She also manages an in-house technology awareness database. She edits the *Annual Review of OCLC Research* and a quarterly newsletter for internal distribution. She is responsible for the Distinguished Seminar Series and moderates the annual Office of Research retreat. She is responsible for designing and implementing systems to control the publications produced by Office of Research staff. She is secretary for the Central Ohio chapter of the American Society for Information Science (COASIS).

**Thomas L. Terrall,** Senior Programmer/Analyst, Office of Research

Ph.D., Physics, The Ohio State University

Dr. Terrall is system administrator for the Apollo computer network. He develops image processing software and applies the X Windows System to support research projects. Previously, he developed software for engineering and manufacturing and conducted research in nuclear theory.

**Kurosh Valanejad,** Technical Assistant, Office of Research

B.S., Computer Science, University of Oklahoma

Mr. Valanejad assists the Library of Congress Class M Subset project and provides technical support for the HyperBach hypermedia reference system.

**David R. VanVoorhis,** Research Assistant, Office of Research

Associate Degree, Electronic Engineering, Hocking Technical College, Nelsonville, Ohio; pursuing a B.S. degree in electrical engineering at The Ohio State University

Mr. VanVoorhis provides programming support and code conversion for various research projects. Previously, he was a communications engineering technician for American Electric Power Corporation and an electronics technician at Ohio University.

**Diane Vizine-Goetz,** Research Scientist, Office of Research

B.A., English; M.L.S., Information and Library Studies, State University of New York at Buffalo; Ph.D., Information and Library Science, Case Western Reserve University, Cleveland, Ohio

Dr. Vizine-Goetz is co-investigator on a joint project for research into increasing the accessibility of the Library of Congress subject headings. The project is supported by the Council on Library Resources, The University of Michigan, and OCLC. She also conducts research on database quality and library cataloging and classification.

**Stuart L. Weibel,** Research Scientist, Office of Research

B.S., Biology, Ohio Dominican College, Columbus, Ohio; Ph.D., Pharmacology, The Ohio State University

Dr. Weibel is project manager for the Automated Document Structure Analysis and Connectionist Models projects and is interested in the application of expert systems to information systems. His other research interests include automated cataloging, optical character recognition, document markup standards, and the application of neural network models to classification and pattern recognition. Previously, he was chairman of the Biology Department and taught Biology at Ohio Dominican College. His computer experience includes development of computer assisted instruction and automated laboratory data collection and analysis.

**Dan N. Wiebe,** Research Assistant, Office of Research

B.A., Computer Science, Ohio Wesleyan University, Delaware, Ohio; pursuing a master's degree at The Ohio State University

Mr. Wiebe conducts research on the SMART experimental information retrieval system.


Bonneau, John R., with D. J. Gardner-Bonneau. n.d. Developing a voice application from scratch: a tale from the trenches. *Journal of the American Voice Input/Output Society*. In press.

Dillon, Martin. n.d. Barriers to the use of research ideas in the design of real systems. In *Online Public Access to Library Files Conference: Influencing the System Designer, Bath, England, September 12–15, 1987*. Forthcoming.

Dillon, Martin, et al. n.d. F-TAS: a full-text access system. In *National Online Meeting Proceedings—1988*. Forthcoming.

Dillon, Martin, B. Mortland, and A. Repka. 1987. D: a design language for retrieval system interfaces. In *National Online Meeting Proceedings—1987, May 5–7, New York*, comp. M. E. Williams and T. H. Hogan, *111–18*. Medford, N.J.: Learned Information, Inc.

Drone, Jeanette M. 1988. Hyperbach: a prototype hypermedia reference system. *OCLC Newsletter* No. 173 (May–June):11–12.

Gordon, Martha J., and M. Dillon. 1988. Factors affecting interface design for full-text retrieval. In *Recherche d'Informations Assistee par Ordinateur (RIAO) Conference on User-Oriented Context-based Text and Image Handling, Massachusetts Institute of Technology, March 21-24*, 607–75. Imprimerie Interne CID.

O'Neill, Edward T., and D. Vizine-Goetz. n.d. Bibliographic relationships: implications for the function of the catalog. In *Proceedings of the Conference on the Conceptual Foundations of Cataloging, University of California, Los Angeles, February 14–15, 1987*, ed. E. Svenonius. Forthcoming.

O'Neill, Edward T., and D. Vizine-Goetz. 1988. Quality control in online databases. *Annual Review of Information Science and Technology*. Forthcoming.

O'Neill, Edward T., with N. Sanders and S. Weibel. n.d. Automated collection analysis using the OCLC and RLG bibliographic databases. *College & Research Libraries*. In Press.

Prabha, Chandra G., and D. Rice. n.d. Assumptions about information seeking behavior in nonfiction books: their importance to full text systems. In *ASIS '88, Proceedings of the 51st ASIS Annual Meeting*. Forthcoming.

Prabha, Chandra G., D. Rice, and D. Cameron. 1988. *Nonfiction Book Use by Academic Library Users*. Research report no. OCLC/OR/RR-88/1. Dublin, Ohio: OCLC.

Prabha, Chandra G., J. Bunge, and D. Rice. 1988. Some aspects of nonfiction book use: a study of public library patrons. *Library & Information Science Research* 10:177–93.

Svenonius, Elaine, and M. Molto. n.d. Automatic derivation of name access points in cataloging. *Journal of the American Society for Information Science*. Forthcoming.

Vizine-Goetz, Diane, and K. Markey. n.d. Subject access literature—1987. *Library Resources and Technical Services*. Forthcoming.

Weibel, Stuart L., M. Oskins, and D. Vizine-Goetz. 1987. Automated title page cataloging. In *ASIS '87, Proceedings of the 50th ASIS Annual Meeting, Boston, Mass., October 4–8*, ed. Ching-chih Chen, 234–38. Medford, N.J.: Learned Information, Inc.

_____. n.d. Automated title page cataloging. *Information Processing and Management*. Forthcoming.

## Presentations

Dillon, Martin. 1987a. "Artificial Intelligence and Expert Systems Applications in Information and Bibliographic Control." Presented at Midwest Federation of Library Associations, Fifth Quadrennial Conference, November 5, Indianapolis, Indiana.

_____. 1987b. "Research Projects at OCLC and the Future of the Library." Presented at the University of Maryland ACN/IRCN Libraries Retreat, December 8, Berkeley Springs, West Virginia.

_____. 1988a. "Current Research Emphasis at OCLC's Office of Research." Presented to the Nebraska Library Commission, February 26, University of Nebraska, Lincoln; Health Sciences OCLC User Group Annual Meeting, April 15, OCLC.

_____. 1988b. "Linguistic Analysis and Library Automation." SIG/LAN panel presentation, ASIS Mid-year Meeting, May 17, Ann Arbor, Michigan.

_____. 1988c. "Structuring Texts for Interaction and Retrieval." Presented at the Study Group on the Structure of Electronic Text Seminar: "The Coming of Age of Electronic Text," May 24, Carnegie Mellon University, Pittsburgh.

Drone, Jeanette M. 1987. "Library of Congress Class M Subset Project." Presented at the Southeast Music Library Association Chapter Meeting, October 2, University of Georgia, Athens.

Drone, Jeanette M., and John A. Whisler. 1987a. "Music and Sound Recording Tagging." Workshop sponsored by the Illinois OCLC User's Group and the Illinois State Library, November 17, Springfield, Illinois; November 19, Chicago.

_____. 1987b. "Uniform Titles for Scores and Sound Recordings." Workshop sponsored by the Illinois OCLC User's Group and the Illinois State Library, November 18, Springfield, Illinois; November 20, Chicago.

Drone, Jeanette M., and Mark Crook. 1988. "HyperBach Demonstration." System demonstration presented at the Music Library Association Annual Meeting, February 11, Minneapolis, Minnesota.

Hickey, Thomas B. 1988a. "User-Oriented Content-based Text and Image Handling." Presented at Recherche d'Informations Assistee par Ordinateur (RIAO) Conference on User-Oriented Content-based Text and Image Handling, March 21–24, Massachusetts Institute of Technology, Cambridge.

_____. 1988b. "Using SGML for Indexing and Displaying a Chemical Encyclopedia." Presented at the Study Group on the Structure of Electronic Text Seminar: "The Coming of Age of Electronic Text," May 23–24, Carnegie Mellon University, Pittsburgh.

Lindeman, Martha Gordon. 1988. "User-Oriented System Design." Presented at the Defense Logistics Agency Automated Data Processing Training Conference, May 4, Columbus, Ohio.

Svenonius, Elaine. 1988. "Thesaurus Construction." Presented at the Special Libraries Association Meeting, June 12, Denver, Colorado.

Terrall, Thomas L. 1988. "Introduction to Fourier Series." Ten-week series of lectures presented as part of the Office of Research Classification Seminar, January–March, OCLC.

Weibel, Stuart L. 1987. "Automated Title Page Cataloging." Presented at the 50th ASIS Annual Meeting, October 4–8, Boston.

# Other OCLC Publications

## OCLC Research Report Series

Through the Research Report Series, OCLC publishes research project results. The research reports are published and distributed by OCLC. In addition, the reports are deposited with the Educational Resources Information Center (ERIC), which makes them available in microfiche and paper copy. OCLC supplies research reports upon request while supplies last until the documents are available from ERIC (usually six to nine months after publication). In addition, reports may be requested through the OCLC Interlibrary Loan Subsystem (OCLC Library Institution Symbol OCC).

*How Public Library Patrons Use Nonfiction Books*
by C. Prabha, J. Bunge, and D. Rice
Report Number: OCLC/OPR/RR-87/1; ERIC ED 281 553; 96 pages; OCLC Control Number: 15170301

*Nonfiction Book Use by Academic Library Users*
by C. Prabha, D. Rice, and D. Cameron
Report Number: OCLC/OR/RR-88/1; 60 pages; OCLC Control Number: 17723892

We hope you have enjoyed reading the *Annual Review of OCLC Research*. Please take a few minutes to tell us what you think by returning the prepaid reader response card.

For further information about OCLC research write to:

**Office of Research**
**Online Computer Library Center, Inc.**
**6565 Frantz Road**
**Dublin, Ohio 43017**

For publication price and ordering information write to:

**Documentation Department**
**Online Computer Library Center, Inc.**
**6565 Frantz Road**
**Dublin, Ohio 43017**

## OCLC Library, Information, and Computer Science Series

Four new OCLC monographs focus on issues of major importance to the library community: library planning, the future of the public library, audiovisual material, and the Linked Systems Project.

***Planning in OCLC Member Libraries,*** edited by M. E. L. Jacob, 1988

*Planning in OCLC Member Libraries* tells how various types of libraries—academic, academic research, state, public, and special libraries—plan their futures.

Monograph 9; 6″ × 9″; 144 pages, softbound, $16.50.

***The Future of the Public Library: Conference Proceedings,*** 1988

*The Future of the Public Library: Conference Proceedings* summarizes the "Conference on the Future of the Public Library," held in Dublin, Ohio, March 20-22, 1988.

Monograph 8; 6″ × 9″; 164 pages, softbound, $13.50

***Audiovisual Material Glossary,*** by Nancy B. Olson, 1988

Compiled by Nancy B. Olson, the *Audiovisual Material Glossary* defines unfamiliar terms and resolves inconsistencies in terminology encountered by those handling audiovisual materials.

Monograph 7; 8 1/2″ × 11″; 52 pages, softbound, $8.50.

***The Linked Systems Project: A Networking Tool for Libraries,*** 1988

*The Linked Systems Project* contains 12 papers describing the progress of the project and the underlying concepts and protocols it uses.

Monograph 6; 6″ × 9″; 152 pages, softbound, $13.50

*Editors* Erik Jul, Mary T. Taylor
*Photography and Design* Martin Hawk
*Typesetter* Tammy Seagle
Distinguished Seminar Series reported by Erik Jul

Thank you for your interest in OCLC research. Please take a few minutes to complete and return the prepaid reader survey card. Your opinion counts.

Detach

1. How much of the *Annual Review of OCLC Research* did you read?
   ☐ All ☐ Almost all ☐ Three-fourths
   ☐ Half ☐ Less than half

2. Rank the sections in order of interest to you (1 = most interesting):
   ___ OCLC Project Reports
   ___ LISRG Program Reports
   ___ Programs
   ___ OCLC Publications
   ___ OCLC Staff
   ___ External and Collaborative Research Reports
   ___ Distinguished Seminar Series

3. In general, what did you like best about the *Annual Review?*
   _____
   _____
   _____

4. What did you like least?_____
   _____
   _____
   _____

5. What would you like to see in future issues?_____
   _____
   _____
   _____

6. The coverage of OCLC research was ☐ too detailed ☐ about right ☐ too general.

7. Other comments:_____
   _____
   _____

8. How did you get the Annual Review? ☐ Direct mailing ☐ Routing slip ☐ Colleague ☐ Conference ☐ Other (describe)_____
   _____

9. Please ☐ add ☐ remove my name from the mailing list (please provide complete mailing address):
   _____
   _____
   _____
   _____

Thank you for your interest in OCLC research. Please take a few minutes to complete and return the prepaid reader survey card. Your opinion counts.

Detach

1. How much of the *Annual Review of OCLC Research* did you read?
   ☐ All ☐ Almost all ☐ Three-fourths
   ☐ Half ☐ Less than half

2. Rank the sections in order of interest to you (1 = most interesting):
   ___ OCLC Project Reports
   ___ LISRG Program Reports
   ___ Programs
   ___ OCLC Publications
   ___ OCLC Staff
   ___ External and Collaborative Research Reports
   ___ Distinguished Seminar Series

3. In general, what did you like best about the *Annual Review?*
   _____
   _____
   _____

4. What did you like least?_____
   _____
   _____
   _____

5. What would you like to see in future issues?_____
   _____
   _____
   _____

6. The coverage of OCLC research was ☐ too detailed ☐ about right ☐ too general.

7. Other comments:_____
   _____
   _____

8. How did you get the Annual Review? ☐ Direct mailing ☐ Routing slip ☐ Colleague ☐ Conference ☐ Other (describe)_____
   _____

9. Please ☐ add ☐ remove my name from the mailing list (please provide complete mailing address):
   _____
   _____
   _____
   _____

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

**BUSINESS REPLY MAIL**

FIRST CLASS PERMIT NO. 11 DUBLIN, OHIO

POSTAGE WILL BE PAID BY ADDRESSEE

**Office of Research**
OCLC Online Computer Library Center, Inc.
6565 Frantz Road
Dublin, Ohio 43017-9983

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

**BUSINESS REPLY MAIL**

FIRST CLASS PERMIT NO. 11 DUBLIN, OHIO

POSTAGE WILL BE PAID BY ADDRESSEE

**Office of Research**
OCLC Online Computer Library Center, Inc.
6565 Frantz Road
Dublin, Ohio 43017-9983

OCLC Online Computer Library Center, a non-profit membership organization, is engaged in computer library service and research and makes available computer-based processes, products and services for libraries and other educational organizations, and library users. More than 8,000 libraries contribute to and/or use information in the OCLC Online Union Catalog, the world's largest database of library bibliographic information. OCLC has participating libraries in 26 countries.

The OCLC Office of Research conducts mission-oriented research to provide the library and information science community with theoretical findings and practical applications.

OCLC Online Computer Library Center, Inc.
6565 Frantz Rd.
Dublin, Ohio 43017-0702
(614)764-6000

Pub. No. 601b
8808/3740-10M, FLG